AI Open 1 (2020) 40–56

Contents lists available at ScienceDirect

AI Open

journal homepage: www.keaipublishing.com/en/journals/ai-open

User behavior modeling for Web search evaluation☆

Fan Zhang [a], Yiqun Liu [a,*], Jiaxin Mao [b], Min Zhang [a], Shaoping Ma [a]

[a] Department of Computer Science and Technology, Institute for Artificial Intelligence, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, China
[b] Gaoling School of Artificial Intelligence, Renmin University of China, China

ARTICLE INFO

Keywords:
Web search
Evaluation metric
User model
User behavior

ABSTRACT

Search engines are widely used in our daily life. Batch evaluation of the performance of search systems to their users has always been an essential issue in the field of information retrieval. However, batch evaluation, which usually compares different search systems based on offline collections, cannot directly take the perception of users to the systems into consideration. Recently, substantial studies have focused on proposing effective evaluation metrics that model user behavior to bring human factors in the loop of Web search evaluation. In this survey, we comprehensively review the development of user behavior modeling for Web search evaluation and related works of different model-based evaluation metrics. From the overview of these metrics, we can see how the assumptions and modeling methods of user behavior have evolved with time. We also show the methods to compare the performances of model-based evaluation metrics in terms of modeling user behavior and measuring user satisfaction. Finally, we briefly discuss some potential future research directions in this field.

1. Introduction

Over the past decades, search engines have become one of the main ways for us to acquire information (Croft et al., 2010). In the research and development of search engines, evaluation has always been a major force, which became central to such an extent that new designs and proposals and their evaluation became one (Saracevic, 1995). Offline evaluation (Sanderson, 2010) (also called *test collection evaluation* or *batch evaluation*) and online evaluation (Hofmann et al., 2016) are two main kinds of approaches for Web search evaluation. Different from online evaluation approaches (e.g. A/B testing (Kohavi et al., 2009) and interleaved comparisons (Joachims, 2002), as well as their variants (Li et al., 2015; Schuth et al., 2015)), which are usually based on implicit measurement of realistic users' experience of systems in a production environment, offline evaluation approaches usually compare different search systems based on test collections. The genesis of test collection evaluation of information retrieval systems can be traced back to the work of Cleverdon (1959), which is well known as the Cranfield paradigm and has been widely used in some IR benchmark workshops (e.g. TREC ,[1] CLEF ,[2] and NTCIR [3]) for many years. In the Cranfield framework, a test collection, which consists of a corpus, representative topics (also referred to as queries), and relevance judgments for query-document pairs, is constructed to access effectiveness of different search systems with official evaluation metrics.

In spite of being repeatable and efficient, offline evaluation approaches cannot directly take the perception of users to the systems into consideration (Moffat et al., 2012). To address this issue, recently, substantial studies have focused on proposing effective evaluation metrics that model user behavior to bring human factors in the loop of Web search evaluation. With underlying user models, the evaluation metrics provide a simulation of users of search systems to bridge the gap between offline evaluation and realistic user behavior. To the best of our knowledge, however, little effort has been made to summarize these studies from the perspective of user behavior modeling in a systematic manner by following the development history of model-based evaluation metrics.

In this paper, we attempt to provide a comprehensive overview of the development of user behavior modeling for Web search evaluation and related works of different model-based evaluation metrics. Specifically,

☆ This work is supported by the National Key Research and Development Program of China (2018YFC0831700), Natural Science Foundation of China (Grant No. 61732008, 61532011), Beijing Academy of Artificial Intelligence (BAAI) and Tsinghua University Guoqiang Research Institute.

* Corresponding author.
E-mail address: yiqunliu@tsinghua.edu.cn (Y. Liu).

[1] https://trec.nist.gov/.
[2] http://www.clef-initiative.eu/.
[3] http://research.nii.ac.jp/ntcir/.

https://doi.org/10.1016/j.aiopen.2021.02.003
Received 1 December 2020; Received in revised form 13 February 2021; Accepted 21 February 2021

we first review the history of how model-based evaluation metrics have evolved and introduce a variety of metrics which design different user models to encode different aspects of user behavior. Then we make detailed comparisons of these metrics in terms of three components: the *user behavior space* they define, the *user decision assumption* they propose, and the *user utility function* they choose. In the following section, we also show the methods to compare how model-based evaluation metrics perform in terms of modeling user behavior and measuring user satisfaction, as well as the consistency between these two facets of the metrics. Finally, we briefly discuss some potential future research directions in the field of user behavior modeling for Web search evaluation.

1.1. Related work

Several previous studies are related to our paper.

Robertson (2000) summarized the components and general principles of a traditional experimental framework design of evaluation, as exemplified by TREC, in information retrieval. Sanderson's (Sanderson, 2010) monograph reviewed more recent examinations of the validity of the test collection approach and evaluation measures, as well as outlining trends in current research exploiting query logs and live labs. The main focus of these works was the high value of the long-standing evaluation method, namely test collection based evaluation, for information retrieval research.

Some other surveys focus on a large number of evaluation metrics used to measure the performance of search systems and try to divide these metrics into different categories. Vyas (2016) discussed various types of metrics under the group of *Graded Relevance Based Metrics* and made a comparison of these metrics based on discriminative power. Bama et al. (2015) presented the summary on two categories of metrics that focused on unranked and ranked sets of documents respectively and provided several illustrations to explain these metrics in their work. Sakai's (Sakai, 2013a) lecture also covered these two categories of "traditional" metrics, which he referred to as *Set Retrieval Metrics* and *Ranked Retrieval Metrics*. Going a step further, besides "traditional" metrics, he also defined and discussed representative "advanced" metrics designed for handling diversity, multi-query sessions, and summarization and question answering systems that go beyond the ranked-list paradigm. The above surveys introduce a wide variety of evaluation metrics in information retrieval mainly based on the properties, such as rank and scale of relevance, of the metrics, rather than the perspective of user behavior modeling.

Recently, some frameworks have been developed to formulate evaluation metrics with corresponding user models. Carterette (2011) organized and described different choices to compose model-based evaluation metrics. He argued that metrics are composed from three underlying models:

- a *browsing model* that describes how a user interacts with the Search Engine Result Page (SERP);
- a *document utility model* that describes how a user derives utility from individual relevant results in the SERP;
- a *utility accumulation model* that describes how a user accumulates utility in the course of browsing.

Different from Carterette, Moffat et al. (2013) explained the metrics with three different but interrelated ways and established the C/W/L framework:

- conditional (**C**) probabilities of a user continuing to read past each rank;
- weights (**W**) on each document rank;
- probabilities of a user leaving (**L**) at a given rank.

Both the above frameworks attempted to explore the connection between evaluation metrics and user models, so as to unified different evaluation metrics according to their properties. Inspired by these frameworks, this paper is also pursuing this line of research. Nevertheless, our work differs from these previous studies in the following folds:

- We systematically sorts out substantial existing model-based evaluation metrics by following their development history so that we can see how the assumptions and modeling methods of user behavior evolved with time.
- We break up the construction of model-based evaluation metrics into three components and summarize the differences in these components among the existing metrics.
- This survey covers some metrics proposed in recent studies, which are not discussed in previous frameworks. In addition, we show the methods of comparing the performance of the metrics in terms of modeling user behavior and measuring user satisfaction, as well as the consistency between these two facets of the metrics.

Note that some related topics are not discussed in detail in this survey. Before the blossoming of resources in the Web, IR evaluation metrics were traditionally developed for set retrieval. These metrics, such as precision, recall, and F-measure, are not connected with user models. Later, there are some widely-used metrics that can be explained with implicit user models. For example, DCG (Järvelin and Kekäläinen, 2002) was proposed to handle the position bias and graded relevance assessments. Carterette (2011) claimed that it is an exemplar of an alternative user model: a user picks a stopping rank k, and then derives utility from all of the relevant documents from ranks 1 through k, with a stopping probability of $1/\log_2(k+1) - 1/\log_2(k+2)$. Similarly, the corresponding user model of average precision (AP) was described by Robertson (2008). It suggests that different users stop scanning the result list at different relevant documents and the stopping probability distribution is assumed to be uniform across all relevant documents. In this paper, however, we mainly focus on those metrics which explicitly encode user models, like rank biased precision (RBP) (Moffat and Zobel, 2008). Another topic is evaluation for diversified search. Considering the trade-off between relevance and diversity, α-nDCG (Clarke et al., 2008) may be the first metric for diversified search evaluation. Further, based on per-intent relevance assessments, Agrawal et al. (2009) proposed to apply the intent-aware (IA) approach to traditional evaluation metrics for diversified search evaluation. For example, ERR-IA, which is an intent-aware version of ERR (Chapelle et al., 2009), may be the most popular of the IA metrics. Diversified search metrics like α-nDCG and ERR-IA are also not covered in this survey. Finally, recent studies have focused more on the development of evaluation metrics for multi-query sessions. In the main body of this survey, we only review the evaluation metrics for individual queries. The session based adaptations of metrics are left for discussion about future research directions.

The remainder of this paper is organized as follows. In Section 2, we review the development history of model-based evaluation metrics. Then, we make a comparison of these metrics in terms of three components in Section 3. Next, in Section 4, we show the methods of measuring the performance of the metrics in two facets. Finally, we conclude the field of user behavior modeling for Web search evaluation with a discussion about future directions in Section 5.

2. Development of model-based evaluation metrics

In this section, we review a lot of representative papers related to model-based evaluation metrics by following their development history. Fig. 1 visually presents an impact graph among these papers. If a paper is heavily based on or inspired by another paper, there is an impact arrow from the former to the latter.

In the impact graph, the position of a paper in the horizontal direction indicates its publication time, while papers with the same height in the vertical direction use similar approaches to model user behavior. For example, the paper of EBU (#4), Click Model-Based Metrics (#8), and

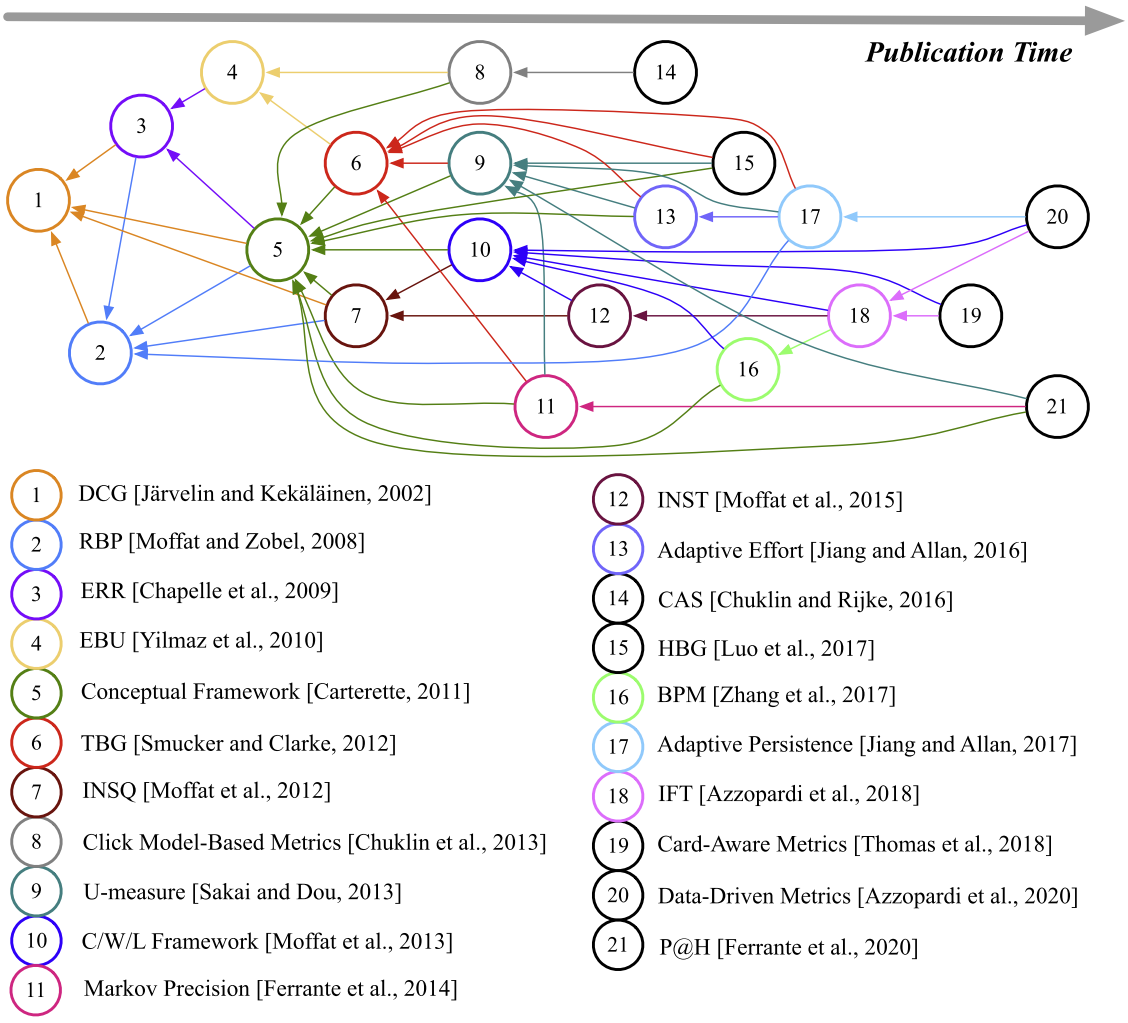

Fig. 1. Impact graph among representative papers related to model-based evaluation metrics. The numbers of the papers are sorted by the time when they were published.

CAS (#14) are all based on click models. In contrast, the paper of Markov Precision (#11) and P@H (#21) both apply Markov Chains to construct user models and evaluation metrics. We can also find that some papers in the center of this impact graph, e.g. Carterette's Conceptual Framework (#5) which attempted to formulate a unified framework to explain existing metrics with user models, have played a significant role in the development of this research area.

In the remaining part of this section, we will introduce the main idea of each paper one by one, and show how they define a user model for search evaluation. Note that these papers may also borrow ideas from some other papers which are not shown in figure. A part of them will be mentioned when we introduce these representative papers in the following.

2.1. Discounted cumulated gain (DCG)

Although the user model of DCG (Järvelin and Kekäläinen, 2002) is only implicitly encoded by its log-based discount which was argued not grounded, we would like to start from this metric given its great popularity and tremendous influence on later metrics. Considering graded relevance judgments, the authors proposed to compute the cumulative gain users obtain from the retrieved results for search performance evaluation. Furthermore, they introduced a log-based discount function to reduce the gain of documents by their ranks with an assumption that documents at greater ranks will be less likely examined due to time, effort, or previous information. The notion of gain and the discount applied to the gain also motivated the design of many future metrics.

2.2. Rank-biased precision (RBP)

RBP (Moffat and Zobel, 2008) could be regarded as the first metric which explicitly defined a user model. It was initially introduced to address the limitations of average precision (AP), which was criticized for not connecting with a reasonable user model. The expression of RBP is:

$$RBP = (1-p)\sum_i r_i p^{i-1} \tag{1}$$

where r_i is the relevance or gain of the document at rank i. In the user model assumed by RBP, users always examine the first result and keep making only one decision, to view the next result with fixed *persistence* (probability) p or not, until they decide to stop. Therefore, this user model is, in fact, a user stopping model with quite simple strategies for stopping. Following RBP, some later metrics, for example ERR which will be introduced next, also focus on the user stopping model, while considering more sophisticated strategies.

2.3. Expected Reciprocal Rank (ERR)

Different from DCG and RBP, which assume that the examination or stopping decision of users is only based on rank positions, ERR (Chapelle et al., 2009) takes the relevance of results into consideration. ERR, derived from a cascade user model (Craswell et al., 2008), assumes that the probability $\mathscr{R}_i$ that users stop after examining a document depends on how they are satisfied with this document, which is directly affected by the relevance r_i of the document with the following function:

$$\mathscr{R}_i = \frac{2^{r_i} - 1}{2^{r_{max}}}, r_i \in \{0, \ldots, r_{max}\} \tag{2}$$

Given a ranking list, we can compute the probability that users stop at rank k with Equation (2):

$$P_k = \left(\prod_{i=1}^{k-1}(1 - \mathscr{R}_i)\right)\mathscr{R}_k \tag{3}$$

Then ERR is defined by $\sum_{k=1}^{n}\frac{1}{k}P_k$, which intuitively indicates where its name *Expected Reciprocal Rank* comes from. As discussed by Chapelle et al. ERR can be seen as a special case of *Normalized Cumulative Utility* (NCU) (Sakai et al., 2008), by choosing the stopping probability P_k and the utility $1/k$.

2.4. Expected Browsing Utility (EBU)

EBU (Yilmaz et al., 2010) also follows the assumption that the probability of leaving depends on the relevance of documents like ERR. However, it takes a further step, by distinguishing between the clicked documents and those that are only scanned. Inspired by related work in the field of click model, specifically, the *Dynamic Bayesian Network* (DBN) click model (Chapelle and Zhang, 2009), the user model of EBU can be tuned by observations from previous click logs. The model can be described as a sequence of actions and decisions of users:

The users always examine the snippet of the first document. Each time they have examined (the snippet of) a document d_i, they decide to click this document for more information or not based on the attractiveness of the snippet, with a probability of $P(C_i|E_i)$. Turpin et al. (2009) suggested that this probability is strongly correlated with the relevance r_i of the document, which means that $P(C_i|E_i) \approx P(C_i|r_i)$. Then, in accordance with the decision whether the users click the document or not, they will leave the search process with a probability of $P(L_i|C_i, E_i)$ or $P(L_i|\overline{C}_i, E_i)$, respectively.

Given this user model, the probability that document d_i can be examined by users can be derived as follows:

$$\begin{aligned} P(E_i) &= P\left(\overline{L}_{i-1}, E_{i-1}\right) \\ &= P\left(\overline{L}_{i-1}|C_{i-1}, E_{i-1}\right)P(C_{i-1}|E_{i-1})P(E_{i-1}) \\ &+ P\left(\overline{L}_{i-1}\middle|\overline{C}_{i-1}, E_{i-1}\right)P\left(\overline{C}_{i-1}\middle|E_{i-1}\right)P(E_{i-1}) \end{aligned} \tag{4}$$

Note that $P(E_1) = 1$, while $P(C_{i-1}|E_{i-1})$ and $P(\overline{L}_{i-1}|C_{i-1}, E_{i-1})$ can be estimated from click logs using DBN with their approximations

$$P(C_{i-1}|E_{i-1}) \approx P(C_{i-1}|r_{i-1}), P\left(\overline{L}_{i-1}|C_{i-1}, E_{i-1}\right) \approx P\left(\overline{L}_{i-1}\middle|r_{i-1}\right) \tag{5}$$

And $P(\overline{L}_{i-1}|\overline{C}_{i-1}, E_{i-1})$ can be regarded as the *persistence* of users when skipping over a document, similar to that in RBP. Therefore, we can estimate $P(E_i)$ according to Equation (4). Finally, the metric corresponding to the user model can be defined as *Expected Browsing Utility* with $\sum_i \mathscr{D}_{EBU}(i)r_i$, where $\mathscr{D}_{EBU}(i)$ is the discount for the gain of document d_i, and $\mathscr{D}_{EBU}(i) = P(E_i)P(C_i|r_i)$.

EBU was the first work to enhance evaluation metrics with click logs, which suggested a new way for later studies.

2.5. Carterette's Conceptual Framework

Previous metrics, such as DCG and RBP, are usually expressed as an inner product of a gain vector and a discount vector. However, the original purposes of using discounts were different for DCG and RBP. In DCG, the discount $F(k)$ is more like the probability that document d_k is viewed by users, while RBP's discount $P(k)$ represents the probability that users stop at rank k. Carterette (2011) pointed out the connection between these two explanations of discounts as follows:

$$F(k) = \sum_{i=k}^{n} P(i) \tag{6}$$

which means if users stop at ranks which are greater or equal to k, the document d_k is viewed by users. Based on this connection, nearly all metrics can be unified with a stopping model and formulated as $\sum_k U(k)P(k)$, where $P(k)$ is the stopping probability and $U(k)$ depends on how we measure the utility of users when they stop at rank k. For example, RBP only focuses on the utility of d_k, which can be expressed as "Model 1: Expected Utility", $M_1 : \sum_{k=1}^{n} rel_k P(k)$, in Carterette's Conceptual Framework. In contrast, DCG focuses on the total utility from d_1 to d_k, thus can be expressed as "Model 2: Expected Total Utility", $M_2 : \sum_{k=1}^{n}(\sum_{i=1}^{k} rel_i)P(k)$, in the framework. Given Equation (6), Carterette showed that M_2 can be also formulated as:

$$\sum_{k=1}^{n}\left(\sum_{i=1}^{k} rel_i\right)P(k) = \sum_{k=1}^{n} rel_k\left(\sum_{i=k}^{n} P(i)\right) = \sum_{k=1}^{n} rel_k F(k) \tag{7}$$

which is the original definition of DCG. Similarly, Zhang et al. (2010) also found that many traditional metrics could be expressed by these different ways.

Besides Model 1 and Model 2, Carterette also discussed two more models with different choices of utility function $U(k)$. Model 3 which focused on the effort (or reciprocal effort), including ERR, was called Expected Effort, while Model 4 combining gain and effort, such as AP, was Expected Average Utility. This framework indeed provided a better understanding of evaluation metrics based on user models and had a significant influence on many related studies on this research line. As shown in Fig. 1, a lot of arrows point to this paper.

2.6. Time-Biased Gain (TBG)

The above metrics do not distinguish user behaviors of interacting with different documents and assume that the time (or effort) users spend on each document is same. However, in modern search scenarios where snippets (a.k.a. summaries) are provided with documents in SERP (Tombros and Sanderson, 1998), how likely users are to read the contents of documents is affected by the quality of snippets. Note that EBU also takes the snippets into account. Different from EBU, which considers the click decisions of users, Smucker and Clarke (2012a) proposed to use time to measure the effort users make to examine snippets and further read documents. In addition, they assumed that shorter documents are expected to take less time than longer documents.

Following Carterette's framework (Carterette, 2011), Smucker and Clarke formulated Time-Biased Gain (TBG) (Smucker and Clarke, 2012a) as the accumulative gain discounted over the expected time $T(k)$ to reach rank k:

$$TBG = \frac{1}{\mathcal{N}} \sum_k g_k D(T(k)) \tag{8}$$

In their paper, g_k was defined as the probability of clicking on a relevant summary $P(C=1|R=1)$ multiplied by the probability of judging the document as relevant $P(S=1|R=1)$ for relevant documents, while zero for non-relevant documents. As for the discount function, they defined $D(T(k)) = e^{-T(k)\frac{\ln 2}{h}}$, where h is the time at which half of the initial users have stopped, called *half life* of users. $T(k)$ is composed of the time to examine the snippet and the expected time to read the document as follows:

$$T(k) = \sum_{i=1}^{k-1} T_S + T_D(l_i) P(C=1|R=r_i) \tag{9}$$

where $T_D(l_i)$ is a linear function of the length l_i of document d_i. Note that all the above probabilities, e.g. $P(C|R)$, $P(S|R)$, and time parameters, e.g. T_S, T_D, and h, can be estimated (or calibrated) with realistic user data.

TBG is not the first work to incorporate time for search evaluation. For example, revisiting the metric *Expected Search Length* (ESL) (Cooper, 1968) introduced by Cooper, Dunlop (1997) proposed *Expected Search Duration*, which was also a time-based evaluation method. Nevertheless, TBG combined the time factor with a general framework which had achieved much success before, encapsulating many widely used metrics such as DCG and RBP. Smucker and Clarke also emphasized that TBG was not the only possible route to do this combination, and the user model could be adapted for better approximation through simulation (Smucker and Clarke, 2012b, 2012c).

2.7. Inverse squares (INSQ)

So far the user stopping model of the above metrics only consider the influence of position and relevance of documents on users' stopping decision. To develop a more reasonable user model, Moffat et al. (2012) noticed the influence of users' intention to find useful information on their stopping decision. The assumption is that the probability of a user continuing search having reached rank i is positively related to T, the anticipated number of relevant documents. The continuation probability is:

$$C_{INSQ}(T,i) = \frac{(i+2T-1)^2}{(i+2T)^2} \tag{10}$$

Once we know the continuation probability C_M of a metric, the leaving (or stopping) probability L_M can be easily inferred with

$$L_M(k) = \prod_{i=1}^{k-1} C_M(i)(1 - C_M(k)) \tag{11}$$

Moffat et al. (2013) further explored the relationship between C_M and L_M, as well as W_M, which is the weight of a document contributing to the metric, and proposed a C/W/L framework.

2.8. Click model-based metrics

Recall that EBU was enhanced with the DBN click model (Chapelle and Zhang, 2009), the parameters of which can be learnt from click logs. With the development of click models, a natural idea is to convert more click models like the User Browsing Model (UBM) (Dupret and Piwowarski, 2008) and the Dependent Click Model (DCM) (Guo et al., 2009) into evaluation metrics. In light of this, Chuklin et al. (2013) discussed how to transform a click model into a metric and summarized a series of click model-based metrics. Referring to the expected utility-based and effort-based metrics, Chuklin et al. showed two ways to map a click model to a metric:

$$\begin{aligned} uMetric &= \sum_{k=1}^{N} P(C_k = 1) R_k \\ rrMetric &= \sum_{k=1}^{N} P(S_k = 1|C_k = 1) P(C_k = 1) \frac{1}{k} \end{aligned} \tag{12}$$

where the parameters related to $P(C_k = 1)$ and $P(S_k = 1|C_k = 1)$ could be estimated with different click models.

2.9. U-measure

Inspired by TBG which discounts the gain of documents over time to examine snippets and read documents, Sakai and Dou (2013) proposed a unified U-measure framework based on *trailtexts* to directly handle the issue of nonlinear traversals. A trailtext tt is concatenated by n strings $s_1 s_2 \ldots s_n$, where each string is a snippet or full text of document in traditional search scenario. Given the concept of trailtext, U-measure was expressed as:

$$U = \frac{1}{\mathcal{N}} \sum_{pos=1}^{|tt|} g(pos) D(pos) \tag{13}$$

where pos is defined as the offset position: $pos(s_k) = \sum_{j=1}^{k} |s_j|$. The position-based gain $g(pos(s_k))$ is related to the relevance of s_k, while the discount function follows that in S-measure framework (Sakai et al., 2011) as:

$$D(pos) = \max\left(0, 1 - \frac{pos}{L}\right) \tag{14}$$

L can be regarded as the largest length of text that users are willing to read. With actual user data or user model assumptions, trailtexts can be constructed and evaluated by U-measure.

2.10. C/W/L framework

C/W/L framework (Moffat et al., 2013) is to explain metrics with three interrelated parts of a user model. The core idea behind this framework is still the user stopping model which focuses on stopping decisions of users. Given the assumptions that characterize how users make stopping decisions, we can infer any one of the following three functions:

The first one is the weight function $W(i)$. It can be regarded as the weight of a document d_i contributing to the metric. If we affirm that a document which is useful to users must be examined by users, $W(i)$ can also be considered as the likelihood that d_i is inspected by users. As expressed by previous works, a general formulation of a metric M is:

$$M = \sum_i W_M(i) r(i) \tag{15}$$

The second one is the continuation probability function $C(i)$. It is a conditional probability that users proceed to rank $i+1$ given that they have reached rank i. Note that if d_i has been examined by users and they decide to proceed to next document, d_{i+1} will be examined too. Therefore, we can derive $C(i)$ from $W(i)$:

$$C_M(i) = \frac{W_M(i+1)}{W_M(i)} \tag{16}$$

The last one is the leaving probability function $L(i)$. It equivalently means the probability that d_i is the last one observed by users. The relationship between $L_M(i)$ and $C_M(i)$ has been shown in Equation (11).

The definition of any one of $W_M(i)$, $C_M(i)$, and $L_M(i)$ completely specifies a metric M. Therefore, given this C/W/L framework, the task of designing evaluation metrics has been simplified as defining $W_M(i)$, $C_M(i)$, or $L_M(i)$ based on a user model.

2.11. Markov Precision (MP)

Ferrante et al. (2014) attempted to provide a more general user model in the framework of Markovian Processes. They assumed that users could start from an arbitrary document and move forward or backward to any other documents after spending a random time on the document. Different from U-measure (Sakai and Dou, 2013) which extracted trail-texts from users' search process, Ferrante et al. (2014) exploited Markov Chains to model this process. Let X_i be a random variable on the state space $\mathscr{T} = \{1, \ldots, T\}$ with an initial distribution $\lambda = (\lambda_1, \ldots, \lambda_T)$. $X_i = j$ means that users examine document d_j at step i. The property of Markov Chain guarantees that

$$P[X_{n+1} = j | X_n = i, X_{n-1}, \ldots, X_0] = P[X_{n+1} = j | X_n = i] = p_{i,j} \tag{17}$$

Given the state space, the initial distribution, and the transition matrix, the random variables (X_n) define a time homogenous discrete time Markov Chain. Inspired by TBG (Smucker and Clarke, 2012a), Ferrante et al. (2014) also obtain a continuous-time Markov Chain by assuming that the times (T_n) spent viewing the documents follow exponential distributions:

$$P[T_n \le t] = \begin{cases} 0 & t < 0 \\ 1 - \exp(-\mu t) & t \ge 0 \end{cases} \tag{18}$$

Based on the user model defined by discrete time or continuous-time Markov Chains, two versions of *Markov Precision* are formulated as:

$$\begin{aligned} MP &= \sum_{i \in \mathscr{R}} \pi_i Prec(i) \\ MP\text{cont} &= \sum_{i \in \mathscr{R}} \tilde{\pi}_i Prec(i) \\ \tilde{\pi}_i &= \frac{\pi_i (\mu_i)^{-1}}{\sum_{j \in \mathscr{R}} \pi_j (\mu_j)^{-1}} \end{aligned} \tag{19}$$

where $Prec(i)$ represents the Precision at step i and π is the (unique) *invariant distribution* of the sub-Markov Chain (Y_n) of (X_n) that considers just the visits to the judged relevant documents.

The main contributions of *Markov Precision* are the novel idea to inject user models into metrics with Markov Chains and the unified framework to bridge the gap between rank-based and time-based metrics via using different versions of Markov Chains.

2.12. INST

INST (Moffat et al., 2015) shares a similar idea with INSQ (Moffat et al., 2012). The difference is that INST takes the assumption that the probability of users stopping search increases as they get closer to their goal of finding T relevant documents into account. Given this assumption, Bailey et al. (2015) described the corresponding continuation probability of INST:

$$C(i) = \left(\frac{i + T + T_i - 1}{i + T + T_i} \right)^2 \tag{20}$$

where T_i is the remaining relevance expected still to be gained. Given the definition of $C(i)$, INST can be easily derived from the C/W/L framework (Moffat et al., 2013). The user model of INST is adaptive, which means as relevant documents are identified, the continuation probability of users will change correspondingly.

2.13. Adaptive effort metrics

Similar to TBG (Smucker and Clarke, 2012a), Jiang and Allan (2016a) considered different effort for different documents and modified existing metrics to *Adaptive Effort* metrics. In contrast to TBG and U-measure, which estimate effort based on document length, the adaptive effort metrics only consider the influence of relevance on the effort. In addition, Jiang and Allan (2016a) directly formulated metrics with adaptive effort, rather than embedding the effort into the discount function.

Different from Carterette's framework (Carterette, 2011), Jiang and Allan (2016a) classified existing metrics into two groups:

$$\begin{aligned} M_1 &= \frac{E(\text{gain})}{E(\text{effort})} = \frac{\sum_i P_{\text{examine}}(i) g(i)}{\sum_i P_{\text{examine}}(i) e(i)} \\ M_2 &= E\left(\frac{\text{gain}}{\text{effort}} \right) = \sum_j P_{\text{stop}}(j) \frac{\sum_{i=1}^{j} g(i)}{\sum_{i=1}^{j} e(i)} \end{aligned} \tag{21}$$

The main idea of adaptive effort metrics is to use an adaptive effort $e(i)$ while retaining the other parts of existing metrics. To apply adaptive efforts, the authors provided two ways. The first one is to fix the ratio of effort between relevant and non-relevant documents, $e_{r/nr}$, and let the effort on relevant documents be one unit. The another approach is to estimate effort based on relevance from search logs like TBG:

$$t(rel) = t_{\text{summary}} + P_{\text{click}}(rel) t_{\text{click}}(rel) \tag{22}$$

2.14. Clicks, attention and satisfaction (CAS) model

The CAS model (Chuklin and de Rijke, 2016) is an extension of click mode-based metrics (Chuklin et al., 2013) by incorporating the effects of "good abandonments" (Li et al., 2009) and non-linear SERP layouts. As its name suggests, the CAS model has three components to model user behavior: an attention model, a click model, and a satisfaction model.

First, the attention model describes that users' attention (or examination) on SERP elements are determined by feature vectors of the elements. In their paper, Chuklin and Rijke (Chuklin and de Rijke, 2016) considered 16 features from three groups and applied a logistic regression ε to convert feature vectors $\vec{\phi}_k$ into examination probabilities:

$$P(E_k = 1) = \varepsilon\left(\vec{\phi}_k \right) \tag{23}$$

After examining an element on SERP, users decide to click the corresponding document or not with the click model. The authors used a generalization of the Position-Based Model (PBM) (Chuklin et al., 2015) as follows:

$$\begin{aligned} P(C_k = 1 | E_k = 0) &= 0 \\ P(C_k = 1 | E_k = 1) &= \alpha_{u_k} \end{aligned} \tag{24}$$

where α_{u_k} is the attractiveness probability of the SERP element u_k.

Finally, the satisfaction model assumes that the examined SERP elements and clicked documents contribute to the total utility of users, thus determine user satisfaction. Given this assumption, the total utility is:

$$U = \sum_k P(E_k = 1) u_d\left(\vec{D}_k \right) + P(C_k = 1) u_r\left(\vec{R}_k \right) \tag{25}$$

where $\vec{D}_k$ and $\vec{R}_k$ are vectors of relevance judgments assigned by different assessors for SERP element and full document of result at k, respectively. u_d and u_r are linear functions that convert histogram of the relevance judgments into an utility. Given the likelihood of satisfaction with

$$P(S = 1) = \sigma(\tau_0 + U) = \sigma\left(\tau_0 + \sum_k \varepsilon\left(\vec{\phi}_k \right) \left(u_d\left(\vec{D}_k \right) + \alpha_{u_k} u_r\left(\vec{R}_k \right) \right) \right) \tag{26}$$

we can train the parameters and obtain a utility metric based on the CAS model.

2.15. Height-biased gain (HBG)

Obviously, HBG (Luo et al., 2017) was inspired by TBG (Smucker and Clarke, 2012a). Its central idea is to cumulate discounted gain over *height*, rather than *time*. Similar to *trailtext* of U-measure (Sakai and Dou, 2013), Luo et al. (2017) introduced *user browsing trail*, which is a concatenation of the contents users have viewed. The difference is that user browsing trail can handle non-textual contents, which are very common in results of modern SERPs, especially in mobile search scenario. The underlying user model of HBG is not much different from that of TBG. It assumes that users first examine the snippet of the first result with height h_1^{SP}. Then they decide to click this result or not based on the *Snippet Relevance* (S_1) and *Click Necessity* (N_1). If they click, then they view the landing page of this result with height h_1^{LP}. Otherwise, they move to examine the snippet of the next result. When users feel satisfied or exhausted, they eventually stop their search process. Note that the authors introduced a novel variable called *Click Necessity*, which means how necessary it is to click a result given its presented snippet. With Click Necessity, the click probability is $P(C_i|S_i,N_i) \approx P(C_i|R_i,N_i)$, following the similar approach in EBU. Relevance R_i and Click Necessity N_i can both be annotated by assessors. Given the click probability, we can estimate the expected viewed height (evh_k) of result k in user browsing trail:

$$evh_k = f^{SP}\left(h_k^{SP}\right) + P(C_k|R_k,N_k) f^{LP}\left(h_k^{LP}\right) \tag{27}$$

Then the formulation of HBG is given by:

$$HBG = \frac{1}{\mathscr{N}} \sum_k \int_{start_k}^{start_{k+1}} G_k(h - start_k) D(h) dh \tag{28}$$

where $start_k$ is the height offset of result k in user browsing trail, which was expressed as $\sum_{i=1}^{k-1} evh_i$. G_k is the distribution function of gain on result k, which was assumed to be a uniform distribution on the expected viewed height of snippet and landing page in their paper. $D(h)$ is the discount function over height, for which the authors considered an exponential decay function or inverse Gaussian decay function, calibrated with real observations.

2.16. Bejeweled player model (BPM)

The user model of BPM (Zhang et al., 2017a) also follows C/W/L framework (Moffat et al., 2013), the focus of which is the user stopping model. Inspired by a game *Bejeweled*, Zhang et al. (2017a) imposed constraints with users' expectations for both gain and effort to determine when users stop. They assumed that users stop at rank k only when:

$$\text{Gain}_k \geq \text{E\_Gain}_k \quad \text{OR} \quad \text{Effort}_k \geq \text{E\_Effort}_k \tag{29}$$

where Gain_k and Effort_k are cumulative gain and effort until rank k, while E_Gain_k and E_Effort_k are the gain users expect to obtain and the effort users are willing to make, respectively. The subscript k of E_Gain_k and E_Effort_k indicates that users' expectations may change at each rank because their information need change through the course of interaction (Fuhr, 2008). Considering whether users' expectations are fixed or allowed to change, Zhang et al. (2017a) provided a static version and a dynamic version of BPM. They also compared three utility functions for BPM, which were total gain, reciprocal effort or average utility, similar to Carterette's framework (Carterette, 2011).

2.17. Adaptive Persistence Metrics

Previous metrics usually have a parameter to represent the *persistence* of users, characterizing user behavior of continuing or stopping. For example, the parameter p in RBP is the continuation probability of examining the next result, while the half life h in TBG is the time at which half of the initial users have stopped. Jiang and Allan (2017) argued that users' persistence should be adaptive on various SERPs with different qualities, thus proposing adaptive persistence versions of existing metrics. Therefore, instead of designing novel user models, they focused on construct an adaptive model for persistence parameters of existing user models behind metrics. Specifically, they modeled a persistence parameter s of a metric as a linear model based on the relevance of results on the SERP:

$$s = w_0 + \sum_{i=1}^{n} \sum_{j=0}^{r_{max}} w_{ij}[r_i = j] \tag{30}$$

where $[r_i = j]$ is a binary variable which equals to 1 if the relevance label of result at rank i is grade j. The parameters w_0 and w_{ij} can be trained on eye-tracking data or click logs.

2.18. Information foraging theory-based measure (IFT)

IFT (Azzopardi et al., 2018) was also defined within the C/W/L framework (Moffat et al., 2013) and imposed constraints which considered the influence of users' expectations on their stopping decisions, similar to INST (Moffat et al., 2015) and BPM (Zhang et al., 2017a). However, IFT seemed to be connected with a more grounded theory, Information Foraging Theory (Pirolli and Card, 1999), which models how people search for information with instinctive foraging mechanisms that similarly adopted to find food through evolution. Based on information foraging theory, Azzopardi et al. (2018) suggested two rules affected the user stopping model.

The first one is a *Goal Sensitive* rule which assumes that users' continuation probability decreases as gain accumulates:

$$C1_i = 1 - \left(1 + b_1 e^{(T-\gamma_i)R_1}\right)^{-1} \tag{31}$$

where T is the target (similar to INST), and γ_i is the gain users have obtained. R_1 is a rationality parameter which indicates the extent to which users are affected by the Goal Sensitive rule. b_1 can be regarded as the stopping probability when users have reached their target.

The second rule is called *Rate Sensitive* rule, which restates Charnov's Marginal Value Theorem (Charnov et al.). It assumes that users' continuation probability decreases as the rate of gain (average gain obtained with a unit effort) decreases:

$$C2_i = \left(1 + b_2 e^{\left(A - \frac{\gamma_i}{\kappa_i}\right)R_2}\right)^{-1} \tag{32}$$

where A is the tolerated rate of gain, below which users are likely to stop. κ_i is the effort users have spent so far. R_2 is also a rationality parameter like R_1 and b_2 operates in a similar way as b_1.

Given the above two rules, the continuation probability of IFT can be expressed as $C_i = C1_i \cdot C2_i$, which can be embedded within the C/W/L framework to derive the IFT metric.

2.19. Card-Aware Metrics

Modern search engine result pages usually contain various "card"-like results such as instant answers, maps, and image frames. To handle the cards and better evaluate the effectiveness of SERPs, Thomas et al. (2018) developed "card-aware" variants of existing metrics. Although the CAS model (Chuklin and de Rijke, 2016) was not mentioned in Card-Aware Metrics paper (Thomas et al., 2018), we find that their underlying user models actually share the same idea and take similar user actions and decisions into consideration. Card-Aware Metrics also assume that users can obtain gain from the card (presented element on SERP) or the full document of a result. Given each card-document pair, users first examine the card, and then have four possible decisions: (1) stop after the card with probability $1 - C_{card,i}$; (2 or 3) continue past the card with

probability $C_{card,i}$, click on the full document with probability E_i, and then (2) stop after the document with probability $1 - C_{doc,i}$ or (3) continue past the document to the next card with probability $C_{doc,i}$; (4) continue past the card with probability $C_{card,i}$, not click with probability $1 - E_i$, and continue to the next card.

Given these decisions, Thomas et al. (2018) defined the continuation probability within C/W/L framework (Moffat et al., 2013) as:

$$C_i = C_{card,i}(E_i C_{doc,i} + (1 - E_i)) \tag{33}$$

and the expected gain from the card and document of a result:

$$r_i = r_{card,i} + C_{card,i} E_i r_{doc,i} \tag{34}$$

Then they derived $\overrightarrow{W}$ from $\overrightarrow{C}$ within C/W/L framework and formulated the metric as $\overrightarrow{W} \cdot \overrightarrow{r}$. Note that the labels of $r_{card,i}$ and $r_{doc,i}$ were assigned by assessors. Different from the CAS model, $C_{card,i}$ and $C_{doc,i}$ here were computed with a continuation function defined by the corresponding metric and the click probability E_i was estimated with a map given card labels and ordering constraints. To handle the non-linear layouts of modern SERPs, the authors assumed the reading order adapted from IFT (Azzopardi et al., 2018): users first look the top two cards in the "core" (main sequence), then examine the "right rail", then go back to the remaining cards in the "core".

2.20. Data-Driven Metrics (DDM)

This DDM work (Azzopardi et al., 2020) also followed the C/W/L framework (Moffat et al., 2013), while estimating the continuation function directly from data, rather than relying on pre-defined user behavior assumptions. A similar idea was adopted to estimate adaptive persistences for different SERPs by Jiang and Allan (2017). However, the adaptive persistence was estimated with a linear model based on the relevance of results on the SERP, while DDM proposed a more general data-driven approach to infer the continuation probability with *Maximum Likelihood Estimate*:

$$C(i+1|i,x) = \frac{\sum_{j=i+1}^{\infty} n(u,j,x)}{\sum_{j=i}^{\infty} n(u,j,x)} \tag{35}$$

where $n(u,j,x)$ is the number of users who stop at rank j given x, which can be any information that affects the continuation probability. For example, Azzopardi et al. (2020) investigated the influences of four factors: *Position* (rank of the result), *Focus* (how focused the query on navigational intents, measured by *click ratio*), *Relevance* (relevance label of the result), and *Type* (item type of the result, such as web, ads, video, etc.).

2.21. Precision at stopping time (P@H)

This work (Ferrante and Ferro, 2020) also modeled user browsing behavior with Markov Chains which allows for both forward and backward transitions. Different from *Markov Precision* (Ferrante et al., 2014), which considered the *stationary distribution* of the Markov Chain to derive the metric, *P@H* (Ferrante and Ferro, 2020) focused on the *stopping time*, which is also an important notion in the theory of the stochastic processes. Given the state space $S = \{End\} \cup \mathcal{N}$, the initial distribution $P[X_1 = 1] = 1$, and the transition probability $p_{i,j} = P[X_{n+1} = j | X_n = i]$, the authors defined three major groups of browsing models with a constraint that users can move only between adjacent rank positions:

- *Deterministic Forward Browsing Model* (DFBM) which assumes that users always sequentially examine each result until the last one.
- *Stochastic Forward Browsing Model* (SFBM) which assumes that users may move forward to the next result with probability p_i or stop.
- *Random Walk Browsing Model* (RWBM) which assumes that users may move forward to the next result with probability p_i, or move backward to the previous result with probability q_i, or stop with probability $1 - p_i - q_i$.

Given the above browsing models, the authors defined the stopping time H as the number of steps before a user stop search: $H = H^{End} - 1$, where

$$H^{End} = \inf\{n \geq 1 : X_n = End\} \tag{36}$$

Then the metric $P@H$ was formulated as a random variable:

$$P@H(r) = \frac{1}{f(H)} \sum_{n=1}^{H} g(k(n), r[X_n]) \tag{37}$$

where $f(h)$ is a real positive non decreasing function, which acts as a proxy of effort. $r[X_n]$ is the relevance of the document at step n and $k(n)$ is the number of times users have visited this document up to step n. The gain function g is computed given the visited time k and relevance r of a document

$$g(k,r) = r(1-\lambda)^{k-1} \tag{38}$$

where $\lambda \in [0,1]$ represents the percentage of utility lost at any visit. Note that $P@H$ is a random variable, so the authors also provided three orders to compare different runs.

2.22. Other recent research

Recently, some new research has extended the above model-based evaluation metrics to practical settings. For example, Wicaksono and Moffat (2018) described a methodology for defining the conditional continuation probability $C(i)$ within the C/W/L framework in practice and evaluated the applicability of the approach based on three large search interaction logs from two different sources. They computed empirical estimates $\widehat{C}(i)$ on impression sequences by micro-averaging across users and queries,

$$\widehat{C}(i) = \frac{\sum_{u \in U} \sum_{P \in \mathscr{P}(u)} N(i,P)}{\sum_{u \in U} \sum_{P \in \mathscr{P}(u)} D(i,P)} \tag{39}$$

or by macro-averaging across users,

$$\widehat{C}(i) = \frac{1}{|U'(i)|} \sum_{u \in U'(i)} \frac{\sum_{P \in \mathscr{P}(u)} N(i,P)}{\sum_{P \in \mathscr{P}(u)} D(i,P)} \tag{40}$$

where U is the set of users and $\mathscr{P}(u)$ is the set of impression sequences for user u. $N(i,P)$ and $D(i,P)$ are the respective numerator and denominator contributions of rank position i in impression sequence P, which can be accumulated by different heuristic rules (Wicaksono and Moffat, 2018). With the estimation of the conditional continuation probability, they explored the extent to which the assumptions associated with INST (Moffat et al., 2015) could be supported.

To enable direct comparison between the estimates of model-based evaluation metrics within the C/W/L framework, Azzopardi et al. (2019) provided a common, extensible, open-source tool called "cwl_eval" and examples of how to use it. In its simplest form, cwl_eval takes the same input as trec_eval, but also provides additional options. With the toolkit, it is straightforward to select and configure metrics. The metrics provided by the toolkit include P@k, AP, ERR, RBP, INSQ, INST, DCG, BPM, TBG, U-Measure, and IFT. As a result, each metric reports a series of values:

- Expected Utility per item examined (EU);
- Expected Total Utility (ETU);

- Expected Cost per item examined (EC);
- Expected Total Cost (ETC);
- Expected number of items to be examined i.e expected depth (ED).

By providing the unification of various metrics within the same package,[4] this toolkit promotes a standardized approach to evaluating search effectiveness.

Later, another work (Breuer et al., Schaer) also introduced a Python package called "repro_eval" for search evaluation. Different from cwl_eval, this toolkit focuses on reactive reproducibility studies of system-oriented IR experiments. Specifically, repro_eval can be used to evaluate the *reproducibility* with a reimplemented IR system in combination with the same test collection of the original experiments. If source code for a reference system is not available, repro_eval can support IR researchers to evaluate their reimplemented reference system and gain insight into how similar the original reference system and the reimplemented reference system are. More details of this toolkit can be found in their public GitHub repository.[5]

In addition, Diaz et al. (2020) introduced the concept of *expected exposure* and proposed a general evaluation methodology based on expected exposure, which had connections to existing retrieval metrics. The principle of *equal expected exposure* in this paper argues that "given a fixed information need, no item should receive more or less expected exposure than any other item of the same relevance grade". To measure the deviation from equal expected exposure, the authors computed the squared error between the target exposure $\varepsilon^\star$ and system exposure ε:

$$l(\varepsilon, \varepsilon^\star) = \left|\left|\varepsilon - \varepsilon^\star\right|\right|_2^2 = \left|\left|\varepsilon\right|\right|_2^2 - 2\varepsilon^{\mathrm{T}}\varepsilon^\star + \left|\left|\varepsilon^\star\right|\right|_2^2 \tag{41}$$

where the first term $\left|\left|\varepsilon\right|\right|_2^2$ is *expected exposure disparity* (EE-D), which measures inequity in the distribution of exposure, and the second term $2\varepsilon^{\mathrm{T}}\varepsilon^\star$ is *expected exposure relevance* (EE-R), which measures how much of the exposure is on relevant documents, while the remaining term $\left|\left|\varepsilon^\star\right|\right|_2^2$ is constant for a fixed information need. A system that achieves optimal EE-R may maximize disparity while a system that minimizes EE-D will have very bad expected exposure relevance. Therefore, there is a tradeoff between the disparity (EE-D) and relevance (EE-R), which cannot be captured by traditional metrics. Finally, to compute the target exposure and system exposure, user models of existing metrics are required to describe how each document is exposed to users.

The above recent studies mainly applied model-based evaluation metrics to specific scenarios and did not propose new assumptions about user models and related metrics. Therefore, we only refer to them briefly here. In the next section, we will make a comparison of model-based evaluation metrics shown in Fig. 1 by decomposing them into different components related to user models.

3. Comparison of model-based evaluation metrics

In Section 2, from the past perspective, we have gone through substantial related works of model-based evaluation metrics. Through the introduction of these studies, we can find that some of them have common ideas in certain aspects of user behavior modeling. In this Section, therefore, we attempt to reconsider these metrics and provide overall comparisons of them from the current perspective. Concretely, we decomposed the design process of model-based evaluation metrics into three steps: defining *user behavior space*, proposing *user decision assumptions*, and choosing *user utility function*. Next, we will sequentially analyze the similarities and differences of the representative papers listed in Section 2 in these three steps.

3.1. User behavior space

To design a user model for evaluation, we should first define the user behavior space which is the set of user behaviors considered to characterize what interactions users may have with search systems. For example, we need to consider the order in which users interact with different search results. We should also determine whether users' interactions with the snippets (or cards) on the SERP need to be taken into account.

Table 1 summarizes the user behaviors which are contained in different papers. We divide these user behaviors into four groups: *Interacting Order*, *Interactions with Snippet/Card*, *Interactions with Document*, and *Stopping Point*. Note that the comparison here on behaviors each work contains is based on the description of the framework in the corresponding paper. The practical metrics implemented in the paper may consider a smaller space of user behaviors. For example, evaluating search performance based on the *trailtexts*, U-measure (Sakai and Dou, 2013) takes into account all the user behaviors listed in Table 1. However, the actual user behavior space defined by U-measure varies depending on how the trailtext is constructed. Below we describe how each of the four groups of user behaviors is defined in different user models behind the metrics.

3.1.1. Interacting order

To find useful information, users interact with one search result after another. Supposing that the results are presented in a ranking list,[6] given the positions of the current and next result examined by users, three kinds of transition are defined to determine the interacting order:

- **Forward (FW) Transition.** Users move forward to the next result.
- **Backward (BW) Transition.** Users go backward to the previous result.
- **Random (RD) Transition.** Users randomly jump to any result.

As shown in Table 1, most metrics assume that users can only move forward to the next result. The exceptions are U-measure (Sakai and Dou, 2013), Markov Precision (MP) (Ferrante et al., 2014), and Precision at Stopping Time (P@H) (Ferrante and Ferro, 2020). U-measure is totally based on the trailtext, which is assumed to be "exactly what the user actually read, in the exact order, during an information seeking process" (Sakai and Dou, 2013). Therefore, if the trailtext is constructed with backward and random transitions between results, U-measure can take account of BW and RD besides FW. The basis for MP and P@H is the Markov chain, the transition matrix $(p_{i,j})$ of which defines the probability to pass from result at any rank i to j. However, P@H (Ferrante and Ferro, 2020) add a constraint that users can move only between adjacent rank positions. Consequently, MP can model FW, BW, and RD, while P@H can model only FW and BW.

3.1.2. Interactions with snippet/card

Given the interacting order, when users jump to a result, they will first interact with the snippet (element card on the SERP) of the result. However, these interactions with snippet are overlooked by some metrics. For the remaining works, we focus on the following four kinds of user behavior on snippet they might consider:

- **Examining the Snippet (E_S).** All the metrics that incorporate interactions with snippet include E_S, as it is a prerequisite for obtaining information from the snippet and generating subsequent interactions with the snippet.

[4] https://github.com/ireval/cwl.
[5] https://github.com/irgroup/repro_eval.

[6] Some papers may consider non-linear layouts of results, e.g. IFT (Azzopardi et al., 2018), Card-Aware Metrics (Thomas et al., 2018), but they usually construct a list of results with reading order assumptions.

Table 1
A comparison on user behavior space among different model-based evaluation metrics. FW, BW, and RD mean Forward, Backward, and Random, respectively. E_S, G_S, and Eff_S mean Examining Snippet, Obtaining Gain from Snippet, and Assessing Effort of Examining Snippet, respectively. Similarly, E_D, G_D, and Eff_D mean the same interactions on Document, rather than Snippet. *Cl* means clicking the document after viewing the snippet. Finally, L_S and L_D mean users can leave their search process after examining a snippet or document.

Metrics	Interacting Order (See 3.1.1)			Snippet/Card (See 3.1.2)				Document (See 3.1.3)			Stopping (See 3.1.4)	
	FW	BW	RD	E_S	G_S	Eff_S	*Cl*	E_D	G_D	Eff_D	L_S	L_D
DCG (Järvelin and Kekäläinen, 2002)	✓							✓	✓			✓
RBP (Moffat and Zobel, 2008)	✓							✓	✓			✓
ERR (Chapelle et al., 2009)	✓							✓	✓	✓		✓
EBU (Yilmaz et al., 2010)	✓			✓			✓	✓	✓		✓	✓
Carterette (Carterette, 2011)	✓							✓	✓	✓		✓
TBG (Smucker and Clarke, 2012a)	✓			✓		✓	✓	✓	✓	✓	✓	✓
INSQ (Moffat et al., 2012)	✓							✓	✓			✓
CMBM (Chuklin et al., 2013)	✓			✓			✓	✓	✓	✓	✓	✓
U-measure (Sakai and Dou, 2013)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
C/W/L (Moffat et al., 2013)	✓							✓	✓	✓		✓
MP (Ferrante et al., 2014)	✓	✓	✓					✓	✓	✓		✓
INST (Moffat et al., 2015)	✓							✓	✓			✓
AdapEM (Jiang and Allan, 2016a)	✓			✓		✓	✓	✓	✓	✓	✓	✓
CAS (Chuklin and de Rijke, 2016)	✓			✓	✓		✓	✓	✓		✓	✓
HBG (Luo et al., 2017)	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓
BPM (Zhang et al., 2017a)	✓							✓	✓	✓		✓
AdapPM (Jiang and Allan, 2017)	✓			✓		✓	✓	✓	✓	✓	✓	✓
IFT (Azzopardi et al., 2018)	✓							✓	✓	✓		✓
CAM (Thomas et al., 2018)	✓			✓	✓		✓	✓	✓		✓	✓
DDM (Azzopardi et al., 2020)	✓							✓	✓	✓		✓
P@H (Ferrante and Ferro, 2020)	✓	✓						✓	✓	✓		✓

- **Obtaining Gain from the Snippet (G_S).** This behavior means that users can also get useful information from the snippet/card, which indicates the effects of good abandonments. U-measure (Sakai and Dou, 2013) defines the position-based gain as $g(pos(s_k))$, where s_k could be an arbitrary part of a text, including, of course, the snippet. CAS (Chuklin and de Rijke, 2016) measures the total utility as the sum of utility from both snippets and full documents, where the utility of snippet is converted from a histogram of the relevance labels assigned by the raters. HBG (Luo et al., 2017) assumes that the gain from the snippet accounts for 40% of the total gain of a result, which is based on Lorigo et al.'s finding that users usually spend 40% of their time looking at the snippet (Lorigo et al., 2008). Card-Aware Metrics (CAM) (Thomas et al., 2018) explicitly denotes the gain from a card at rank i as $r_{card,i}$.
- **Assessing Effort of examining the Snippet (Eff_S).** The effort users make to examine the snippet is also considered by some works. TBG (Smucker and Clarke, 2012a) discounts the gain of a result over time spent on all the snippets and documents, and the authors calibrated time to assess the snippet as $T_S = 4.4s$ from their data. Inspired by TBG, U-measure (Sakai and Dou, 2013) and HBG (Luo et al., 2017) measure the effort users spend on the snippet in terms of its length and height, respectively. Similar to TBG, Adaptive Effort Metrics (Jiang and Allan, 2016a) adopts an approach to measure effort by the amount of time required to examine a result and uses the same value of T_S. Adaptive Persistence Metrics (Jiang and Allan, 2017) also takes T_S into consideration while estimating the value of T_S on two different datasets.
- **Clicking the document after viewing the snippet (*Cl*).** We need to clarify that metrics which ignore interactions with the snippet also contain click behavior. However, they assume that users will definitely click and find useful information from the document. In our paper, *Cl* actually indicates whether it is possible for users to skip the result without click. Similar to E_S, *Cl* is also included in all the metrics with snippet interactions. It is an action which connects interactions from the snippet to the document.

3.1.3. Interactions with document

After interacting with the snippet, and clicking the full document of the result, users will **Examining the Document (E_D).** For those metrics that overlook snippet interactions, their user models assume that users will directly examining the document once they reach the result. Therefore, E_D is included in all the user models behind evaluation metrics. The same applies to G_D, which denotes **Obtaining Gain form the Document**. It is not surprising, as "classic" evaluation metrics are constructed based on the gain of the documents and the probability that each document is examined by users, such as DCG and RBP. However, the effort of examining the document has not been focused in each paper. These papers usually assume that the amount of effort users spend on each document is same, and has no impact on user behavior or search performance.

3.1.4. Stopping Point

The last group of user behaviors is the stopping behavior. There are two different Stopping Points: **Leaving search after interactions with the Snippet (L_S)**, and **Leaving search after interactions with the Document (L_D)**. The latter one is a basic user behavior that each user model of metrics contain, while the former one is only incorporated in the metrics which simultaneously take interactions with Snippet into account.

3.2. User decision assumptions

Given the definition of the user behavior space of a user model, we should propose a few assumptions about user decisions to describe how users switch between different behaviors and change their search states. In this survey, we mainly focus on two decisions of users: *Stopping Decision* and *Clicking Decision*. Table 2 summarizes the factors that influence users' stopping and clicking decisions among different papers.

3.2.1. Factors influencing stopping decision

Judging whether users stop after interacting with a result is one of the prime concerns for model-based evaluation metrics since it is highly related to the probability that a result will be viewed by users and its contribution to the search performance. As discussed in Section 3.1, users may leave search after interactions with the snippet or document. In this part, we only focus on the stopping decision after viewing a document,

Table 2
A comparison on factors that influence users' current stopping and clicking decisions among different model-based evaluation metrics. Pos, Rel, C_G, C_Eff, and Exps mean Position, Relevance of the current result, Cumulative Gain, Cumulative Effort, and users' Expectations, respectively. Rel_D, Rel_S, and Ness mean Relevance of the Document, Relevance of the Snippet, and Click Necessity of the result, respectively.

Metrics	Factors Influencing Stopping Decision (See 3.2.1)						… Clicking (See 3.2.2)		
	Pos	Rel	C_G	C_Eff	Exps	Others	Rel_D	Rel_S	Ness
DCG (Järvelin and Kekäläinen, 2002)	✓								
RBP (Moffat and Zobel, 2008)									
ERR (Chapelle et al., 2009)		✓							
EBU (Yilmaz et al., 2010)		✓					✓		
Carterette (Carterette, 2011)	✓	✓							
TBG (Smucker and Clarke, 2012a)				✓			✓		
INSQ (Moffat et al., 2012)	✓				✓				
CMBM (Chuklin et al., 2013)	✓	✓					✓		
U-measure (Sakai and Dou, 2013)				✓			✓		
C/W/L (Moffat et al., 2013)	✓	✓	✓		✓				
MP (Ferrante et al., 2014)	✓								
INST (Moffat et al., 2015)	✓		✓		✓				
AdapEM (Jiang and Allan, 2016a)	✓	✓		✓			✓		
CAS (Chuklin and de Rijke, 2016)	✓					✓	✓		
HBG (Luo et al., 2017)				✓			✓		✓
BPM (Zhang et al., 2017a)			✓	✓	✓				
AdapPM (Jiang and Allan, 2017)	✓	✓			✓		✓		
IFT (Azzopardi et al., 2018)			✓	✓	✓	✓			
CAM (Thomas et al., 2018)	✓	✓	✓		✓	✓		✓	
DDM (Azzopardi et al., 2020)	✓	✓				✓			
P@H (Ferrante and Ferro, 2020)	✓	✓							

and analyze different factors considered to influence the decision. The factors influencing the stopping decision after viewing the snippet have been little discussed in the related papers, and we think they can be inspired from factors discussed here.

Note that the stopping decision at current rank i can be formulated as the conditional probability $P(L_i|E_i)$[7] that users stop after viewing the result at rank i. It can also be regarded as the opposite of the conditional continuation probability C_i defined in the C/W/L framework (Moffat et al., 2013), which means $P(L_i|E_i) = 1 - C_i$. Based on the above definition, the factors that influence the conditional probability $P(L_i|E_i)$ or C_i are discussed separately as follows.

As shown in Table 2, *Position* (**Pos**) and *Relevance* (**Rel**) are the two most frequently considered factors which have an influence on the stopping decision. For example, derived from the log-based discount of DCG, which is a function of position i, the conditional stopping probability after viewing result i is also a function of i. Different from DCG, RBP assumes that the conditional stopping probability is fixed as a *persistence* parameter p, so the stopping decision of RBP is not affected by any factors. Some metrics take the influence of relevance of the current result into account. For example, ERR assumes that the conditional stopping probability depends on how users are satisfied with the result, which is affected by its relevance: $P(L_i|E_i) = \mathscr{R}_i = \frac{2^{r_i}-1}{2^{r_{max}}}$. Modeling from another perspective, EBU estimates the conditional stopping probability defined in DBN click model as an approximation based on relevance: $P(L_i|Cl_i, E_i) \approx P(L_i|r_i)$. Some later studies, either by proposing a unified framework (e.g. Carterette's framework (Carterette, 2011), C/W/L framework (Moffat et al., 2013), Click-Model Based Metrics (Chuklin et al., 2013), P@H (Ferrante and Ferro, 2020)) or variants of existing metrics (e.g. Adaptive Effort Metrics (Jiang and Allan, 2016a), Adaptive Persistence Metrics (Jiang and Allan, 2017), Data-Driven Metrics (Azzopardi et al., 2020)), have attempted to take the influence of both position and relevance on the conditional stopping probability into account.

Users' *Cumulative Effort* (**C_Eff**) is also considered as factors influencing stopping decisions. TBG (Smucker and Clarke, 2012a) may be the first work to incorporate the influence of effort (measured by time). Given its exponential decay over time, the conditional stopping probability is only affected by the amount of time users have spent up to the current result, i.e. the cumulative effort. Similar to TBG, U-measure (Sakai and Dou, 2013) and HBG (Luo et al., 2017) also focus on the influence of cumulative effort, while measuring it with the offset position in the trailtext and the current height in the user browsing trail, respectively. Adaptive Effort Metrics (Jiang and Allan, 2016a) also include variants of TBG and U-measure, thus having considered the influence of cumulative effort. In addition, cumulative effort is also considered in BPM (Zhang et al., 2017a) and IFT (Azzopardi et al., 2018), given that these metrics incorporate some other factors besides cumulative effort, we will discuss them later.

Users' *Cumulative Gain* (**C_G**) and *Expectations* (**Exps**) are usually considered together in these papers. INSQ (Moffat et al., 2012) introduces users' anticipated number of relevant documents T, and assumes that the conditional continuation probability is positively related to T: $C_{INSQ}(T, i) = \frac{(i+2T-1)^2}{(i+2T)^2}$. Taking a further step, INST (Moffat et al., 2015) suggests that the conditional continuation probability is also affected by the gain (CG_i) that has been accumulated so far: $C(i) = \left(\frac{i+T+T_i-1}{i+T+T_i}\right)^2$, where T_i is the remaining gain to be obtained beyond position i, and $T_i = T - CG_i$. The same idea is shared by C/W/L framework (Moffat et al., 2013) and Card-Aware Metrics (Thomas et al., 2018), which also consider the influence of both users' cumulative gain and expectations of gain. BPM (Zhang et al., 2017a) and IFT (Azzopardi et al., 2018) are recent works which consider both cumulative gain and cumulative effort, as well as users' expectations related to gain and effort. BPM assumes that users make a decision to stop search when either the cumulative gain reaches the expected gain, or the cumulative effort reaches the expected effort. Differently, IFT adopts the Goal Sensitive and Rate Sensitive rules based on Information Foraging Theory to model the conditional continuation probability. Besides, Adaptive Persistence Metrics (Jiang and Allan, 2017) also takes into account the influence of users' expectations to some extent. The authors suggest that the persistence parameter, which determines whether users stop or not, should be adaptively based on all the results on the SERP. That is to say, users have an expectation of quality for the whole SERP, which will affect the

[7] In the remainder of this paper, we call $P(L_i|E_i)$ as the *conditional stopping probability* to distinguish it from the leaving probability L_i, which is the probability that users leave search at rank i, not given the previous search process.

stopping decision of users in the process of search.

In addition to the factors discussed above, some works have also take into account other factors (**Others**) that may have an influence on stopping decision. CAS model (Chuklin and de Rijke, 2016) applies a logistic regression to convert feature vectors into examination probabilities. The features include some factors such as SERP item type and geometry features. Similarly, different cards (SERP elements) are distinguished in IFT (Azzopardi et al., 2018), Card-Aware Metrics (Thomas et al., 2018), and Data-Driven Metrics (Azzopardi et al., 2020), which have an influence on users' stopping decision.

3.2.2. Factors influencing clicking decision

Since the clicking behavior is only considered in metrics that take users' interactions with snippet into account, the factors influencing clicking decision are discussed in these papers. The most common factor considered by the papers to influence the clicking decision is the *relevance of document* (Rel_D). Most metrics share a same assumption that whether users click a document is affected by the attractiveness of the SERP element, which is strongly correlated with the relevance r_i of the document (Turpin et al., 2009):

$$P(Cl_i|E_i) = a_{d_i} \approx a(r_i) = P(Cl_i|r_i) \tag{42}$$

There are also two studies which consider other factors beyond the relevance of document. HBG (Luo et al., 2017) explicitly defines a novel concept, *Click Necessity* (**Ness**), to model the cases in which the SERP element can directly fulfill users' information needs. Therefore, the authors expresse the probability of click as $P(Cl_i|R_i,N_i)$, which considers the influence of both relevance and click necessity. Given card labels, Card-Aware Metrics (Thomas et al., 2018) estimates the clicking probability as a mapping from the *relevance of card* (Rel_S).

3.3. User utility function

Given the user behavior space and user decision assumptions, we have been able to describe the complete process of how users interact with search results. The final step to derive a model-based evaluation metric is to choose a proper utility function as a proxy to measure the performance of the systems or the satisfaction of users. Table 3 summarizes the utility functions adopted in different model-based evaluation metrics.

Gain and effort, also referred to as benefit and cost (Azzopardi and Zuccon, 2016), are the pivotal variables of utility functions. Most metrics focus on the gain, since to find useful information is the primary goal of users. Given the weight function $W_M(i)$ of a metric M within the C/W/L framework, Moffat et al. (2013) generalized the formulation of a metric M as: $M = \sum_{i=1}^{\infty} W_i r_i$, where r_i is the relevance (gain) of the result at rank i and W_i denotes $W_M(i)$ for simplicity in what follows. The metrics with the above formulation are called as *Weighted Precision metrics* in C/W/L framework (Moffat et al., 2013), while referred to as *Expected Utility metrics* by Azzopardi et al. (2018). However, we notice that metrics with this formulation can be explained with different utility functions, given two interpretations of the weight function W_i:

1. In the first, W_i can be regarded as the probability that result i is examined by users. Under this explanation, W_i is usually not normalized, which means $\sum_{i=1}^{\infty} W_i \neq 1$. For example, most metrics assume that users always examine the first result, i.e. $W_1 = 1$, which results in that the sum of W_i does not converge. In C/W/L framework, L_i is defined as the probability that result i is the last one observed by users (also can be regarded as the stopping probability at rank i). Given the first interpretation of W_i and the assumption of $W_i = 1$, we can derive L_i as: $L_i = W_i - W_{i+1}$. The intuition is that stopping at rank i means examining result i while not examining result $i+1$. Therefore, the "Weighted Precision" metric M is

$$\begin{aligned} M &= \sum_{i=1}^{\infty} W_i r_i \\ &= \sum_{i=1}^{\infty} W_i r_i + \left[\sum_{i=1}^{\infty} W_i \left(\sum_{j=1}^{i-1} r_j \right) - \sum_{i=1}^{\infty} W_{i+1} \left(\sum_{j=1}^{i} r_j \right) \right] \\ &= \sum_{i=1}^{\infty} W_i \left(\sum_{j=1}^{i} r_j \right) - \sum_{i=1}^{\infty} W_{i+1} \left(\sum_{j=1}^{i} r_j \right) \\ &= \sum_{i=1}^{\infty} (W_i - W_{i+1}) \left(\sum_{j=1}^{i} r_j \right) \\ &= \sum_{i=1}^{\infty} L_i \left(\sum_{j=1}^{i} r_j \right) \end{aligned} \tag{43}$$

where L_i is the stopping probability at rank i and $\sum_{j=1}^{i} r_j$ is the cumulative gain users have obtained up to rank i. The formulation of $\sum_{i=1}^{\infty} L_i (\sum_{j=1}^{i} r_j)$ is called *Expected Total Utility* metrics by Azzopardi et al. (2018). In our paper, to facilitate the comparison with other utility functions, we denote their utility function as E (C_Gain), which means the **Expected**

Table 3
A comparison on formulation of user utility function among different model-based evaluation metrics.

Metrics	User Utility Function					
	E (Gain)	E (C_Gain)	E ($\frac{1}{\text{C\_Effort}}$)	E ($\frac{\text{C\_Gain}}{\text{C\_Effort}}$)	$\frac{\text{E(C\_Gain)}}{\text{E(C\_Effort)}}$	Others
DCG (Järvelin and Kekäläinen, 2002)		✓				
RBP (Moffat and Zobel, 2008)	✓					
ERR (Chapelle et al., 2009)			✓			
EBU (Yilmaz et al., 2010)		✓				
Carterette (Carterette, 2011)	✓	✓	✓	✓		
TBG (Smucker and Clarke, 2012a)		✓				
INSQ (Moffat et al., 2012)	✓					
CMBM (Chuklin et al., 2013)	✓		✓			
U-measure (Sakai and Dou, 2013)		✓				
C/W/L (Moffat et al., 2013)	✓	✓	✓			
MP (Ferrante et al., 2014)				✓		
INST (Moffat et al., 2015)					✓	
AdapEM (Jiang and Allan, 2016a)				✓	✓	
CAS (Chuklin and de Rijke, 2016)		✓				
HBG (Luo et al., 2017)		✓				
BPM (Zhang et al., 2017a)		✓	✓	✓		
AdapPM (Jiang and Allan, 2017)	✓	✓	✓			
IFT (Azzopardi et al., 2018)	✓	✓				
CAM (Thomas et al., 2018)	✓					
DDM (Azzopardi et al., 2020)	✓	✓	✓			
P@H (Ferrante and Ferro, 2020)				✓	✓	✓

Cumulative Gain obtained by users. Many metrics, such as DCG, EBU, TBG, U-measure, CAS, and HBG only adopt this utility function E (C_Gain) by explaining the weight W_i as the examining probability of result i.

2. The alternative interpretation of W_i is the expected proportion of attention users give to result i. It can also be regarded as the contribution of result i on search evaluation. This explanation usually indicates that W_i is normalized with $\sum_{i=1}^{\infty} W_i = 1$. Given this constraint, the stopping probability is $L_i = \frac{W_i - W_{i+1}}{W_1}$. Then the "Weighted Precision" metric M can be rewritten as

$$M = \sum_{i=1}^{\infty}(W_i - W_{i+1})\left(\sum_{j=1}^{i} r_j\right) = W_1 \sum_{i=1}^{\infty} L_i \left(\sum_{j=1}^{i} r_j\right) \tag{44}$$

where the first equality holds by Equation (43). Note that the expected number of results examined by users is given by

$$\sum_{i=1}^{\infty} iL_i = \sum_{i=1}^{\infty} i\frac{W_i - W_{i+1}}{W_1} = \frac{\sum_{i=1}^{\infty} i(W_i - W_{i+1})}{W_1} = \frac{\sum_i W_i}{W_1} = \frac{1}{W_1} \tag{45}$$

If we assume that the effort on each result is one, then $\sum_{i=1}^{\infty} iL_i$ can also be regarded as the **Expected Cumulative Effort** spent by users. Therefore, the metric M can be expressed as

$$M = W_1 \sum_{i=1}^{\infty} L_i \left(\sum_{j=1}^{i} r_j\right) = \frac{\sum_{i=1}^{\infty} L_i \left(\sum_{j=1}^{i} r_j\right)}{\sum_{i=1}^{\infty} iL_i} = \frac{\mathrm{E(C\_Gain)}}{\mathrm{E(C\_Effort)}} \tag{46}$$

That is to say, if W_i is normalized and the effort is same for each result, the "Weighted Precision" metrics in C/W/L framework adopt a utility function which is equivalent to Expected Cumulative Gain divided by Expected Cumulative Effort, $\frac{\mathrm{E(C\_Gain)}}{\mathrm{E(C\_Effort)}}$. In Table 3, if a metric is formulated as $\sum_i W_i r_i$ with a normalized W_i, we denote its utility function as E (Gain), following the name of "**Expected Utility (Gain)**", although it may be also equivalent to the formulation $\frac{\mathrm{E(C\_Gain)}}{\mathrm{E(C\_Effort)}}$ on certain conditions. Metrics like RBP, INSQ, and Card-Aware Metrics are included in this class. On the contrary, if a metric explicitly express its formulation as **Expected Cumulative Gain divided by Expected Cumulative Effort**, we denote its utility function as $\frac{\mathrm{E(C\_Gain)}}{\mathrm{E(C\_Effort)}}$. For example, Moffat et al. (2015) provided an algorithm to compute INST for a document ranking, where the score was divided by the sum of weight *sumW*. Jiang and Allan (2016a) explicitly summarized a group of metrics implemented as $\frac{\mathrm{E(C\_Gain)}}{\mathrm{E(C\_Effort)}}$ and considered different effort for different results. Ferrante and Ferro (2020) also incorporated this formulation given the definition of order $\preccurlyeq_2$ in their paper.

In addition, two other utility functions are also used in some metrics.

- One is the **Reciprocal Cumulative Effort.** For example, ERR can be expressed as $\sum_k \frac{1}{k} P_k$, where P_k is the stopping probability at rank k. The reciprocal rank $1/k$ can be regarded as 1/C_Effort, so the utility function of ERR can be expressed as E ($\frac{1}{\mathrm{C\_Effort}}$). Similar formulation is also considered by BPM. Note that some works propose a unified framework (e.g. Carterette's framework, C/W/L framework, and Click-Model Based Metrics) or variants of existing metrics (e.g. Adaptive Persistence Metrics and Data-Driven Metrics). They include the utility function of E ($\frac{1}{\mathrm{C\_Effort}}$) and some other utility functions given different metrics corresponding to the framework or variant.
- The other one is the **Expected Average Gain**, which can be formulated as E ($\frac{\mathrm{C\_Gain}}{\mathrm{C\_Effort}}$). The metrics with this utility function focus on the average gain obtained by users when they stop at rank k, usually with a formulation as $\sum_k \frac{\mathrm{C\_Gain}_k}{\mathrm{C\_Effort}_k} P_k$, where P_k is the stopping probability at rank k. For example, Markov Precision is based on the precision at each rank k, which can be considered as the average gain at rank k. Ferrante and Ferro (2020) presented the difference between the utility functions E ($\frac{\mathrm{C\_Gain}}{\mathrm{C\_Effort}}$) and $\frac{\mathrm{E(C\_Gain)}}{\mathrm{E(C\_Effort)}}$ very clearly. They proposed a metric P@H of a system run r based on Markov Chain:

$$P@H(r) = \frac{1}{f(H)} \sum_{n=1}^{H} g(k(n), r[X_n]) \tag{47}$$

where $G(H, r) = \sum_{n=1}^{H} g(k(n), r[X_n])$ can be regarded as the Cumulative Gain and $f(H)$ can be regarded as the Cumulative Effort. Note that P@H is a random variable. To compare different runs, an order among random objects is required to define. The first order of P@H between different runs r and s suggested by Ferrante and Ferro is formulated as follows:

$$P@H(r) \underline{\prec}_1 P@H(s) \Leftrightarrow \mathbb{E}\left[\frac{G(H,r)}{f(H)}\right] \leq \mathbb{E}\left[\frac{G(H,s)}{f(H)}\right] \tag{48}$$

The order $\underline{\prec}_1$ indicates that the metric chooses $\mathrm{E}_{\frac{\mathrm{(C\_Gain)}}{\mathrm{(C\_Effort)}}}$ as the utility function. Slightly modified from the order $\underline{\prec}_1$, the second order is

$$P@H(r) \underline{\prec}_2 P@H(s) \Leftrightarrow \frac{\mathbb{E}[G(H,r)]}{\mathbb{E}[f(H)]} \leq \frac{\mathbb{E}[G(H,s)]}{\mathbb{E}[f(H)]} \tag{49}$$

The order $\underline{\prec}_2$ indicates that the metric chooses $\frac{\mathrm{E(C\_Gain)}}{\mathrm{E(C\_Effort)}}$ as the utility function. Besides the above two orders, Ferrante and Ferro (2020) also defined a stochastic order based on the notion of stochastic dominance (Hadar and Russell, 1969):

$$P@H(r) \underline{\prec}_3 P@H(s) \Leftrightarrow \mathbb{P}\left[\frac{G(H,r)}{f(H)} > x\right] \leq \mathbb{P}\left[\frac{G(H,s)}{f(H)} > x\right], \forall x \in \mathbb{R} \tag{50}$$

The order $\underline{\prec}_3$ is a partial order, indicating a stochastic utility function which has not been considered in previous works.

4. Measuring the performance of model-based evaluation metrics

Previous sections have introduced the development of model-based evaluation metrics and the differences between the metrics in terms of *user behavior space*, *user decision assumption*, and *user utility function*. Then the question naturally arises: how to compare the performance of different metrics? Or, how to "evaluate" different evaluation metrics? The related field of research to answer this question is the meta-evaluation of evaluation metrics.

One of the primary motivations for developing model-based evaluation metrics is to connect the metrics with more realistic underlying user models, thus better reflecting real user perception of search performance. Therefore, in this section, we mainly focus on two facets of model-based evaluation metrics related to this motivation:

- How to measure the performance of model-based evaluation metrics on providing more realistic user model?
- How to measure the performance of model-based evaluation metrics on reflecting more realistic user satisfaction?

Besides these two facets, there are some other aspects of metrics considered by meta-evaluation methods. For example, Sakai (2006) proposed *discriminative power* to measure the extent to which evaluation metrics can discriminate among different systems. Another widely-used approach is to calculate the correlation coefficient (e.g. Kendall's τ) between the system orderings determined by different metrics. Considering the characteristic that top positions are more important in IR, Yilmaz et al. (2008) and Webber et al. (2010) introduced an AP-based correlation coefficient called τ_{ap} and *Rank Biased Overlap* (RBO), respectively. In addition, Lu et al. (2016) also reviewed other criteria applied to compare metrics, such as judgment cost, coverage, inversions, and volatility matrix. *Judgment cost* is the cost of a metric in terms of judgment effort. For

example, shallow metrics and steeply top-weighted metrics are cheaper to evaluate to a given level of score fidelity than deep metrics. The *coverage ratio* is the coverage of the set of significantly different system pairs discriminated by the new metric relative to the set of pairs that are found to be separable by the reference (baseline) metric. It measures the extent to which the new metric is able to confirm the relationships that are significantly noted by the reference metric. Considering that the new metric might also deliver the opposite outcome to the reference metric, the *inversion ratio* is computed similar to the *coverage ratio*, while accumulating the pairs for which the new metric provides an opposite result to the reference metric. A *volatility matrix* is a means of visualizing the strength of the relationship between two versions with different truncated depth of a metric. For example, suppose that a metric at depth k_1 and k_2 results in different orderings of systems S as $\sigma_{k_1}(S)$ and $\sigma_{k_2}(S)$, respectively. Then the corresponding component of the matrix will be $corr(\sigma_{k_1}(S),\sigma_{k_2}(S))$, where $corr()$ is one of the rank correlation coefficients such as Kendall's τ. Given the introductions of the above meta-evaluation methods, we can obviously find that they have little relevance with user behavior modeling, thus are not discussed in our paper.

Next, we will introduce the methods used in previous works to measure the performance of model-based evaluation metrics on providing realistic user model and reflecting realistic user satisfaction.

4.1. Performance on providing realistic user model

As Carterette (2011) said, "one possible definition of a good measure is one that more closely models user behavior". Therefore, an intuitive way to evaluate the metrics is comparing how well the user behavior characterized (predicted) by different metrics fits the user behavior observed from real data. The core of the user model behind a metric is the user decision assumptions we have discussed in Section 3.2, so many works focus on the error between users' actual stopping/clicking decisions and those assumed by user models of the metrics.

Given a metric, we can derive the stopping probability of the user at each rank i, as well as the viewing probability that result i is examined by the user. RBP, for example, assumes that the user stops at rank i with a probability of $L_i = p^{i-1}(1-p)$, and the result i is examined by the user with a probability of $V_i = p^{i-1}$, where p is the *persistence* parameter. However, we usually have no way of knowing where the user actually views and stops from the user log. To address this issue, different methods were adopted to estimate users' viewing and stopping behaviors.

The first method is to collect fixations of users captured by eye-tracking devices. If eye fixations are observed on a result, we can assume that the result has been examined by the user. Given this assumption, Jiang and Allan (2017) performed 10-fold cross validations and reported the mean negative log likelihood in predicting eye fixations with different browsing models of metrics. Compared to eye fixations, click signals are cheaper and more easily accessible, thus used by more studies to infer viewing behavior of users. A simple method is to assume that the user stops at the last clicked result at rank i and has examined all the results before. Based on this assumption, Carterette (2011) made a comparison between empirical distribution of stopping ranks and distributions corresponding to different user models. A more sophisticated estimation of viewing probability V_i of user u for query q was suggested by Wicaksono et al. (2019):

$$\widehat{V}_i(u,q) = \begin{cases} 1 & i \le DC(u,q) \\ e^{-(i-DC(u,q))/g(K(u,q))} & otherwise \end{cases} \tag{51}$$

where $K(u,q) = w_0 + w_1 \cdot DC(u,q) + w_2 \cdot NC(u,q)$; $DC(u,q)$ is the deepest rank position clicked and $NC(u,q)$ is the number of distinct items clicked; $g(x) = \ln(1+e^x)$ is a softplus function; and (w_0, w_1, w_2) are parameters estimated from the data. Given the estimation of V_i, Wicaksono et al. (Wicaksono and Moffat, 2020) estimated the empirical distributions of C_i, W_i, and L_i within C/W/L framework, and calculated the *mean squared error* (MSE) between model-generated distributions and empirical distributions to measure the extent to which the model-predicted behavior matches observed user behavior.

Another point to note is that the user stopping probability of some metrics may also depend on factors beyond position (discussed in Section 3.2.1), which means the model-generated distributions of C_i, W_i, and L_i may vary across different queries. To compare them with the empirical distributions over all queries, Wicaksono et al. (Wicaksono and Moffat, 2020) and Zhang et al. (2020a) averaged the values of model-generated distributions across all rank lists in the dataset.

Besides the stopping decision, some metrics which incorporate interactions with the snippet also focus on how well the user models of metrics predict click events. Click prediction is an essential task for click models (Wang et al., 2016), with widely used metrics such as perplexity (Dupret and Piwowarski, 2008) or likelihood of clicks. Similarly, some metrics based on click models, such as EBU (Yilmaz et al., 2010) and CAS (Chuklin and de Rijke, 2016), measured the performance of user models behind metrics in terms of the Root Mean Squared Error (RMSE) between the probabilities of clicks, and the log-likelihood of clicks, respectively.

4.2. Performance on reflecting realistic user satisfaction

Use satisfaction can be understood as "the fulfillment of a specified desire or goal" in information retrieval (Kelly, 2009), which measures users' actual feelings about a system. Recent researches (Al-Maskari et al., 2007; Huffman and Hochster, 2007; Jiang et al., 2015) have revealed a surprisingly strong correlation between IR evaluation metrics and user satisfaction so that it can also be regarded as the golden standard in search performance evaluation. In view of this, a number of studies compare different metrics by measuring their agreement with user satisfaction.

The most direct way is to calculate the correlation coefficient between the scores of metrics and user satisfaction across different queries. Based on some user study datasets which collected self-reported satisfaction of users, many researchers have computed the Pearson correlation coefficient (Pearson's r) to measure the relationship between metric scores and user-reported satisfaction (Jiang and Allan, 2016a; Chuklin and de Rijke, 2016; Zhang et al., 2017a; Jiang and Allan, 2017). Given that user satisfaction is usually collected as a Likert scale item, Zhang et al. (2020a) reported the results of Spearman correlation coefficient (Spearman's ρ), which is a nonparametric measure of rank correlation. Besides calculating the correlation coefficient, Chen et al. (2017) also conduct concordance test (Sakai, 2013b) to compare different metrics in terms of the proportion of pairs where the metric and user satisfaction have an agreement.

Another method is to employ a surrogate for user satisfaction with implicit feedbacks. For example, Chapelle et al. (2009) considered various click metrics which might be good measurements of user satisfaction, including:

- number of clicks in a session (QCTR);
- binary click indicator (UCTR);
- maximum, mean and minimum reciprocal ranks of the clicks (Max, Mean, Min RR);
- search success where relevant documents are clicked (SS);
- the number of clicks divided by the position of the lowest click (PLC).

Besides comparing the correlation between these online click metrics and offline model-based evaluation metrics, Chuklin et al. (2013) also assessed the agreement between offline metrics and the evaluation outcomes of the Team-Draft Interleaving (TDI) method (Radlinski et al., 2008). Different from previous work, Thomas et al. (2018) used query reformulation as a proxy measure of user satisfaction (Hassan et al., 2013). They argued that "reformulation is an indicator of failure in regard to the query which is reformulated" and computed the reformulation rate and success rate of queries in logs.

In addition, user preference collected with the side-by-side method (Thomas and Hawking, 2006) can also be regarded as the golden standard to reflect real experiences of users. Therefore, some studies may compare evaluation metrics in terms of the agreement with user preferences (Sanderson et al., 2010; Luo et al., 2017).

4.3. Consistency between two facets

Existing works have delved into the relationship between user behavior and user satisfaction. For example, Fox et al. (2005) published the most influential early work in user satisfaction prediction and found a strong correlation between search activity and user satisfaction. Kim et al. (2014) estimated distributions of click dwell time for both satisfied and dissatisfied clicks for different click segments and demonstrated that click-level satisfaction has a strong correlation with click dwell time. Liu et al. (2015) investigated the relationship between mouse movement patterns and user satisfaction. They built a model based on the motifs to predict user satisfaction. The above studies have focused on predicting user satisfaction with user behavior information, not considering how well model-based evaluation metrics, as the bridge between user behavior and user satisfaction, perform on the consistency between these two facets.

To shed light on this issue, Wicaksono and Moffat (2020) presented how the user model accuracy and user satisfaction vary depending on the parameters of metrics respectively. The accuracy of user model was measured by mean squared error between the model-generated and observed distributions of C_i, W_i, and L_i functions within C/W/L framework. Note that the user satisfaction is the session-level rating, so they aggregated the query-level scores of evaluation metrics to a session score for comparison. Similarly, Zhang et al. (2020a) investigated how metrics calibrated with user satisfaction performed on measuring query-level satisfaction and found a strong consistency for model-based evaluation metrics between the user model and user satisfaction. They also explored the stability of parameters of metrics and data requirements for the calibration. Their findings indicated that calibrating evaluation metrics with a small scale of user behavior data is an effective and feasible way for search evaluation.

5. Conclusion

5.1. Future directions

There are also several ongoing or future related research topics which are not discussed in our paper but well worth the effort to explore:

- **Session-Based User Models for Evaluation.** Recently, session search evaluation has been paid more attention since realistic search tasks usually involve multiple queries and multi-round search interactions (e.g. exploratory search (White and Roth, 2009)). Existing works mainly evaluate session search in two ways: to aggregate the metric scores of individual queries composing a session (Jiang and Allan, 2016b; Liu et al., 2018, 2019); or to extend query-level metrics to session level based on similar user behavior assumptions (e.g. sDCG (Järvelin et al., 2008) and sRBP (Lipani et al., 2019)). However, these works have not proposed a novel user model specifically designed for session search. Zhang et al. (2020b) have made an attempt to incorporate the recency effect into the utility accumulation model to better evaluate the performance for a whole session. The design of more practical session-based user models still remains to be an important open question.
- **Theoretical User Models for Evaluation.** Most underlying user models of metrics introduced in this paper are based on our intuitive assumptions about user behavior or simple patterns directly observed from user logs. Recently, researchers have been working on proposing theoretically sound user models, including the Game-Theoretic Framework for IR (Zhai, 2016), the interactive Probability Ranking Principle (Fuhr, 2008), the Interface Card Model (Zhang and Zhai, 2015), etc. Promising progress has also been made in combining theories from other disciplines to model user behavior for search evaluation. For example, Azzopardi suggested that the Production Theory () in microeconomics could be used to model the search process and proposed economics IR models (Azzopardi, 2011), which provided the theoretical basis for economic measures like IFT (Azzopardi et al., 2018). Inspired by the recency effect (Baddeley, 1968) found in psychology, Zhang et al. (2020b) modified the utility accumulation model of users and proposed Recency-aware Session-based Metrics.
- **Simulation of User Models for Evaluation.** Another interesting area of research is evaluating search systems based on simulation of users. It is an essential step to evaluate interactive IR systems with reproducible experiments where dynamic test collections are required, because systems may update their retrieved results in response to user actions (e.g. query reformations) dynamically. Carterette et al. (2015) have already made an attempt to provide dynamic test collections by simulating user interactions and propose some novel methods for simulation. Zhang et al. (2017b) have taken a step forward to propose a general formal framework for interactive IR evaluation based on the user simulator and reward/cost of the simulator.

6. Summary

This paper has reviewed a wide range of model-based evaluation metrics and shown how these metrics model user behavior for Web search evaluation. Through a historical overview of the development of model-based evaluation metrics and a comprehensive comparison of the related papers in terms of different components including user behavior space, user decision assumption, and user utility function, we can see some trends in the field of user behavior modeling for Web search evaluation in recent years:

- More kinds of user behavior have been taken into account to characterize user interactions with search systems. For example, given the presentation of various card-like results on modern search engine result pages, user interactions with the snippet or card element have been incorporated into the design of user model for recent metrics.
- More sophisticated user decision strategies have been adopted to describe the actions of users in different search states. For example, considering the influence of factors such as users' expectations of gain and effort on their stopping decision, recent adaptive metrics allow users to stop with different probabilities depending on what they have encountered so far.
- Effort has played a more important role in both modeling user behavior and measuring user utility. For example, some metrics estimate the effort with time spent during the search process and incorporate it into decay function or utility function.

We have also briefly described some methods for meta-evaluating the model-based evaluation metrics in terms of their performances on modeling user behavior and reflecting user satisfaction, as well as the consistency between these two facets in this paper. We believe this research will help researchers quickly grasp some of the well-established ideas and approaches involved in user behavior modeling for Web search evaluation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Agrawal, R., Gollapudi, S., Halverson, A., Ieong, S., 2009. Diversifying search results. In: Proceedings of the Second ACM International Conference on Web Search and Data Mining, pp. 5–14.

Al-Maskari, A., Sanderson, M., Clough, P., 2007. The relationship between ir effectiveness measures and user satisfaction. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 773–774.

Azzopardi, L., 2011. The economics in interactive information retrieval. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 15–24.

Azzopardi, L., Zuccon, G., 2016. An analysis of the cost and benefit of search interactions. In: Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval, pp. 59–68.

Azzopardi, L., Thomas, P., Craswell, N., 2018. Measuring the utility of search engine result pages: an information foraging based measure. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 605–614.

Azzopardi, L., Thomas, P., Moffat, A., cwl_eval, 2019. An evaluation tool for information retrieval. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1321–1324.

Azzopardi, L., White, R.W., Thomas, P., Craswell, N., 2020. Data-driven evaluation metrics for heterogeneous search engine result pages. In: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, pp. 213–222.

Baddeley, A., 1968. Prior recall of newly learned items and the recency effect in free recall. Canadian Journal of Psychology/Revue canadienne de psychologie 22 (3), 157.

Bailey, P., Moffat, A., Scholer, F., Thomas, P., 2015. User variability and ir system evaluation. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 625–634.

Bama, S.S., Ahmed, M., Saravanan, A., 2015. A survey on performance evaluation measures for information retrieval system. International Research Journal of Engineering and Technology 2 (2), 1015–1020.

T. Breuer, N. Ferro, M. Maistro, P. Schaer, Repro Eval: A python Interface to Reproducibility Measures of System-Oriented Ir Experiments.

Carterette, B., 2011. System effectiveness, user models, and user utility: a conceptual framework for investigation. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 903–912.

Carterette, B., Bah, A., Zengin, M., 2015. Dynamic test collections for retrieval evaluation. In: Proceedings of the 2015 International Conference on the Theory of Information Retrieval, pp. 91–100.

Chapelle, O., Zhang, Y., 2009. A dynamic bayesian network click model for web search ranking. In: Proceedings of the 18th International Conference on World Wide Web, pp. 1–10.

Chapelle, O., Metlzer, D., Zhang, Y., Grinspan, P., 2009. Expected reciprocal rank for graded relevance. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 621–630.

E. L. Charnov, et al., Optimal Foraging, the Marginal Value Theorem.

Chen, Y., Zhou, K., Liu, Y., Zhang, M., Ma, S., 2017. Meta-evaluation of online and offline web search evaluation metrics. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 15–24.

Chuklin, A., de Rijke, M., 2016. Incorporating clicks, attention and satisfaction into a search engine result page evaluation model. In: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, pp. 175–184.

Chuklin, A., Serdyukov, P., De Rijke, M., 2013. Click model-based information retrieval metrics. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 493–502.

Chuklin, A., Markov, I., Rijke, M.d., 2015. Click models for web search, Synthesis lectures on information concepts. retrieval, and services 7 (3), 1–115.

Clarke, C.L., Kolla, M., Cormack, G.V., Vechtomova, O., Ashkan, A., Büttcher, S., MacKinnon, I., 2008. Novelty and diversity in information retrieval evaluation. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 659–666.

Cleverdon, C.W., 1959. The evaluation of systems used in information retrieval. In: Proceedings of the International Conference on Scientific Information, vol. 1. National Academy of Sciences, Washington, DC, pp. 687–698.

Cooper, W.S., 1968. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. Am. Doc. 19 (1), 30–41.

Craswell, N., Zoeter, O., Taylor, M., Ramsey, B., 2008. An experimental comparison of click position-bias models. In: Proceedings of the 2008 International Conference on Web Search and Data Mining, pp. 87–94.

Croft, W.B., Metzler, D., Strohman, T., 2010. Search Engines: Information Retrieval in Practice, vol. 520. Addison-Wesley Reading.

Diaz, F., Mitra, B., Ekstrand, M.D., Biega, A.J., Carterette, B., 2020. Evaluating stochastic rankings with expected exposure. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 275–284.

Dunlop, M.D., 1997. Time, relevance and interaction modelling for information retrieval. In: ACM SIGIR Forum, vol. 31. ACM, New York, NY, USA, pp. 206–213.

Dupret, G.E., Piwowarski, B., 2008. A user browsing model to predict search engine click data from past observations. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 331–338.

Ferrante, M., Ferro, N., 2020. Exploiting stopping time to evaluate accumulated relevance. In: Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval, pp. 169–176.

Ferrante, M., Ferro, N., Maistro, M., 2014. Injecting user models and time into precision via Markov chains. In: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 597–606.

Fox, S., Karnawat, K., Mydland, M., Dumais, S., White, T., 2005. Evaluating implicit measures to improve web search. ACM Trans. Inf. Syst. 23 (2), 147–168.

Fuhr, N., 2008. A probability ranking principle for interactive information retrieval. Inf. Retr. 11 (3), 251–265.

Guo, F., Liu, C., Wang, Y.M., 2009. Efficient multiple-click models in web search. In: Proceedings of the Second Acm International Conference on Web Search and Data Mining, pp. 124–131.

Hadar, J., Russell, W.R., 1969. Rules for ordering uncertain prospects. Am. Econ. Rev. 59 (1), 25–34.

Hassan, A., Shi, X., Craswell, N., Ramsey, B., 2013. Beyond clicks: query reformulation as a predictor of search satisfaction. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 2019–2028.

Hofmann, K., Li, L., Radlinski, F., 2016. Online evaluation for information retrieval. Foundations and Trends in Information Retrieval 10 (1), 1–117.

Huffman, S.B., Hochster, M., 2007. How well does result relevance predict session satisfaction?. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 567–574.

Järvelin, K., Kekäläinen, J., 2002. Cumulated gain-based evaluation of ir techniques. ACM Trans. Inf. Syst. 20 (4), 422–446.

Järvelin, K., Price, S.L., Delcambre, L.M., Nielsen, M.L., 2008. Discounted cumulated gain based evaluation of multiple-query ir sessions. In: European Conference on Information Retrieval. Springer, pp. 4–15.

Jiang, J., Allan, J., 2016a. Adaptive effort for search evaluation metrics. In: European Conference on Information Retrieval. Springer, pp. 187–199.

Jiang, J., Allan, J., 2016b. Correlation between system and user metrics in a session. In: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, pp. 285–288.

Jiang, J., Allan, J., 2017. Adaptive persistence for search effectiveness measures. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 747–756.

Jiang, J., Hassan Awadallah, A., Shi, X., White, R.W., 2015. Understanding and predicting graded search satisfaction. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 57–66.

Joachims, T., 2002. Optimizing search engines using clickthrough data. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 133–142.

Kelly, D., 2009. Methods for evaluating interactive information retrieval systems with users. Foundations and trends in Information Retrieval 3 (12), 1–224.

Kim, Y., Hassan, A., White, R.W., Zitouni, I., 2014. Modeling dwell time to predict click-level satisfaction. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 193–202.

Kohavi, R., Longbotham, R., Sommerfield, D., Henne, R.M., 2009. Controlled experiments on the web: survey and practical guide. Data Min. Knowl. Discov. 18 (1), 140–181.

Li, J., Huffman, S., Tokuda, A., 2009. Good abandonment in mobile and pc internet search. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 43–50.

Li, L., Kim, J.Y., Zitouni, I., 2015. Toward predicting the outcome of an a/b experiment for search relevance. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, pp. 37–46.

Lipani, A., Carterette, B., Yilmaz, E., 2019. From a user model for query sessions to session rank biased precision (srbp). In: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, pp. 109–116.

Liu, Y., Chen, Y., Tang, J., Sun, J., Zhang, M., Ma, S., Zhu, X., 2015. Different users, different opinions: predicting search satisfaction with mouse movement information. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 493–502.

Liu, M., Liu, Y., Mao, J., Luo, C., Ma, S., 2018. Towards designing better session search evaluation metrics. In: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1121–1124.

Liu, M., Mao, J., Liu, Y., Zhang, M., Ma, S., 2019. Investigating cognitive effects in session-level search user satisfaction. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 923–931.

Lorigo, L., Haridasan, M., Brynjarsdóttir, H., Xia, L., Joachims, T., Gay, G., Granka, L., Pellacini, F., Pan, B., 2008. Eye tracking and online search: lessons learned and challenges ahead. J. Am. Soc. Inf. Sci. Technol. 59 (7), 1041–1052.

Lu, X., Moffat, A., Culpepper, J.S., 2016. The effect of pooling and evaluation depth on ir metrics. Information Retrieval Journal 19 (4), 416–445.

Luo, C., Liu, Y., Sakai, T., Zhang, F., Zhang, M., Ma, S., 2017. Evaluating mobile search with height-biased gain. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 435–444.

Moffat, A., Zobel, J., 2008. Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. 27 (1), 1–27.

Moffat, A., Scholer, F., Thomas, P., 2012. Models and metrics: Ir evaluation as a user process. In: Proceedings of the Seventeenth Australasian Document Computing Symposium, pp. 47–54.

Moffat, A., Thomas, P., Scholer, F., 2013. Users versus models: what observation tells us about effectiveness metrics. In: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, pp. 659–668.

Moffat, A., Bailey, P., Scholer, F., Thomas, P., 2015. Inst: an adaptive metric for information retrieval evaluation. In: Proceedings of the 20th Australasian Document Computing Symposium, pp. 1–4.
Pirolli, P., Card, S., 1999. Information foraging. Psychol. Rev. 106 (4), 643.
Radlinski, F., Kurup, M., Joachims, T., 2008. How does clickthrough data reflect retrieval quality?. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 43–52.
Robertson, S., 2000. Evaluation in information retrieval. In: European Summer School on Information Retrieval. Springer, pp. 81–92.
Robertson, S., 2008. A new interpretation of average precision. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 689–690.
Sakai, T., 2006. Evaluating evaluation metrics based on the bootstrap. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 525–532.
Sakai, T., 2013b. Metrics, statistics, tests. In: PROMISE Winter School. Springer, pp. 116–163.
Sakai, T., 2013b. How intuitive are diversified search metrics? concordance test results for the diversity u-measures. In: Asia Information Retrieval Symposium. Springer, pp. 13–24.
Sakai, T., Dou, Z., 2013. Summaries, ranked retrieval and sessions: a unified framework for information access evaluation. In: Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 473–482.
Sakai, T., Robertson, S., Newswatch, I., 2008. Modelling a user population for designing information retrieval metrics. EVIA@ NTCIR.
Sakai, T., Kato, M.P., Song, Y.-I., 2011. Click the search button and be happy: evaluating direct and immediate information access. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management, pp. 621–630.
Sanderson, M., 2010. Test collection based evaluation of information retrieval systems. Foundations and Trends® in Information Retrieval 4 (4), 247–375.
Sanderson, M., Paramita, M.L., Clough, P., Kanoulas, E., 2010. Do user preferences and evaluation measures line up?. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 555–562.
Saracevic, T., 1995. Evaluation of evaluation in information retrieval. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 138–146.
Schuth, A., Hofmann, K., Radlinski, F., 2015. Predicting search satisfaction metrics with interleaved comparisons. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 463–472.
Smucker, M.D., Clarke, C.L., 2012a. Time-based calibration of effectiveness measures. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 95–104.
Smucker, M.D., Clarke, C.L., 2012b. Modeling user variance in time-biased gain. In: Proceedings of the Symposium on Human-Computer Interaction and Information Retrieval, pp. 1–10.
Smucker, M.D., Clarke, C.L., 2012c. Stochastic simulation of time-biased gain. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2040–2044.
Thomas, P., Hawking, D., 2006. Evaluation by comparing result sets in context. In: Proceedings of the 15th ACM International Conference on Information and Knowledge Management, pp. 94–101.
Thomas, P., Moffat, A., Bailey, P., Scholer, F., Craswell, N., 2018. Better effectiveness metrics for serps, cards, and rankings. In: Proceedings of the 23rd Australasian Document Computing Symposium, pp. 1–8.
Tombros, A., Sanderson, M., 1998. Advantages of query biased summaries in information retrieval. In: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2–10.
Turpin, A., Scholer, F., Jarvelin, K., Wu, M., Culpepper, J.S., 2009. Including summaries in system evaluation. In: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 508–515.
Vyas, J.P., 2016. Survey of Graded Relevance Metrics for Information Retrieval.
Wang, C., Liu, Y., Ma, S., 2016. Building a click model: from idea to practice. CAAI Transactions on Intelligence Technology 1 (4), 313–322.
Webber, W., Moffat, A., Zobel, J., 2010. A similarity measure for indefinite rankings. ACM Trans. Inf. Syst. 28 (4), 1–38.
White, R.W., Roth, R.A., 2009. Exploratory search: beyond the query-response paradigm, Synthesis lectures on information concepts. retrieval, and services 1 (1), 1–98.
Wicaksono, A.F., Moffat, A., 2018. Empirical evidence for search effectiveness models. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, pp. 1571–1574.
Wicaksono, A.F., Moffat, A., 2020. Metrics, user models, and satisfaction. In: Proceedings of the 13th International Conference on Web Search and Data Mining, pp. 654–662.
Wicaksono, A.F., Moffat, A., Zobel, J., 2019. Modeling user actions in job search. In: European Conference on Information Retrieval. Springer, pp. 652–664.
Yilmaz, E., Kanoulas, E., Aslam, J.A., 2008. A simple and efficient sampling method for estimating ap and ndcg. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 603–610.
Yilmaz, E., Shokouhi, M., Craswell, N., Robertson, S., 2010. Expected browsing utility for web search evaluation. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp. 1561–1564.
Zhai, C., 2016. Towards a game-theoretic framework for text data retrieval. IEEE Data Eng. Bull. 39 (3), 51–62.
Zhang, Y., Zhai, C., 2015. Information retrieval as card playing: a formal model for optimizing interactive retrieval interface. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 685–694.
Zhang, Y., Park, L.A., Moffat, A., 2010. Click-based evidence for decaying weight distributions in search effectiveness metrics. Inf. Retr. 13 (1), 46–69.
Zhang, F., Liu, Y., Li, X., Zhang, M., Xu, Y., Ma, S., 2017a. Evaluating web search with a bejeweled player model. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 425–434.
Zhang, Y., Liu, X., Zhai, C., 2017b. Information retrieval evaluation as search simulation: a general formal framework for ir evaluation. In: Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval, pp. 193–200.
Zhang, F., Mao, J., Liu, Y., Xie, X., Ma, W., Zhang, M., Ma, S., 2020a. Models versus satisfaction: towards a better understanding of evaluation metrics. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 379–388.
Zhang, F., Mao, J., Liu, Y., Ma, W., Zhang, M., Ma, S., 2020b. Cascade or recency: constructing better evaluation metrics for session search. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 389–398.